A Comprehensive Survey on Knowledge Distillation

Amir M. Mansourian, Rozhan Ahmadi*, Masoud Ghafouri*, Amir Mohammad Babaei*, Elaheh Badali Golezani*,
Zeynab Yasamani Ghamchi*, Vida Ramezanian*, Alireza Taherian*, Kimia Dinashi, Amirali Miri, Shohreh Kasaei
Image Processing Lab, Sharif University of Technology
Email: {amir.mansurian, roz.ahmadi, masoud.ghafouri98, amir.babaei79, elahe.badali, zeynab.yasamani77,
vidaramezanian, alireza.taherian49, kimia.dinashi, amirali.miri79, kasaei}@sharif.edu

Abstract—Deep Neural Networks (DNNs) have achieved notable performance in the fields of computer vision and natural language processing with various applications in both academia and industry. However, with recent advancements in DNNs and transformer models with a tremendous number of parameters, deploying these large models on edge devices causes serious issues such as high runtime and memory consumption. This is especially concerning with the recent large-scale foundation models, Vision-Language Models (VLMs), and Large Language Models (LLMs). Knowledge Distillation (KD) is one of the prominent techniques proposed to address the aforementioned problems using a teacher-student architecture. More specifically, a lightweight student model is trained using additional knowledge from a cumbersome teacher model. In this work, a comprehensive survey of knowledge distillation methods is proposed. This includes reviewing KD from different aspects: distillation sources, distillation schemes, distillation algorithms, distillation by modalities, applications of distillation, and comparison among existing methods. In contrast to most existing surveys, which are either outdated or simply update former surveys, this work proposes a comprehensive survey with a new point of view and representation structure that categorizes and investigates the most recent methods in knowledge distillation. This survey considers various critically important subcategories, including KD for diffusion models, 3D inputs, foundational models, transformers, and LLMs. Furthermore, existing challenges in KD and possible future research directions are discussed. Github page of the project: https://github.com/IPL-Sharif/KD_Survey

Index Terms—Knowledge Distillation, Knowledge Transfer, Teacher-Student Architecture.

I. INTRODUCTION

With the emergence of DNNs, the fields of Computer Vision (CV) and Natural Language Processing (NLP) have been revolutionized, and most tasks in these domains are now solved using DNNs. While a simple ResNet [1] model or BERT [2] can be easily trained on most of existing GPUs, the advent of large models like LLMs and foundational vision models has made training and inference a significant challenge in terms of runtime and memory usage, especially for deployment on edge devices like mobile phones due to their widespread use. Despite their high performance, these large models often have complex architectures and are heavily overparameterized [3], [4]. For example, the weight matrices in LLMs have been shown to be low-rank matrices [5].

To address the issue of large models, various solutions have been proposed, such as efficient network architectures,

Fig. 1. A general teacher-student framework for knowledge distillation.

compression methods, pruning, quantization, low-rank factorization, and knowledge distillation. Efficient network blocks, including MobileNet [6], ShuffleNet [7], and BiSeNet [8] have been introduced in recent years. Pruning methods, a type of compression technique, aim to remove unnecessary layers and parameters from models with minimal impact on performance. Low-rank factorization methods reduce parameters through matrix decomposition. Unlike other methods, KD does not modify the network's layers or parameters. In KD, a lightweight network (student) is trained under the supervision of a more complex model (teacher) with deeper architecture and greater number of parameters. This concept, first proposed by [9] and later popularized by [10] as KD, introduced the idea of training the student network by mimicking the teacher's output distribution as soft labels.

Aside from the large number of parameters, large models also require a significant amount of labeled data for training. Another purpose of KD is knowledge transfer, which can involve transferring knowledge from a source task to a target task that lacks sufficient labeled data. Additionally, in cases where data privacy is a concern, data-free knowledge distillation becomes helpful. This approach addresses the issue by generating synthetic data, eliminating the need to store sensitive data. The key challenges in KD are determining what knowledge to transfer, selecting the appropriate algorithm, and designing the student and teacher architectures. A diagram of a general teacher-student KD method is shown in Figure 1.

With the significant growth in the number of published papers in the field of KD and its wide applications across various tasks and domains, several survey papers have been presented to review KD from different perspectives. The prominent work [11] provides a comprehensive survey on KD,

TABLE I
COMPARISON BETWEEN EXISTING KD SURVEYS AND THIS SURVEY.

Paper	Source	Algorithm	Modality	Application
Gou et al. [11]	Logit, Feature, Similarity	Adversarial, Attention, Cross-modal, Data-free, Graph, Lifelong, Multi-teacher, NAS, Quantized	Image, Text, Speech, Video	Visual Recognition
Wang et al. [12]	Logit, Feature, Similarity	Adversarial, Cross-modal, Data-free, Few-shot, Graph, Incremental, Multi-teacher, Reinforcement	Image, Video	Visual Recognition, Self-supervised Learning
Hu et al. [15]	Logit, Feature, Similarity, Mutual Information	Cross-modal, Federated, Graph, Multi-teacher	Image, Text, Speech, Video	Visual Recognition, Ranking, Generation, Regression
Moslemi et al. [16]	Logit, Feature, Similarity	Adversarial, Attention, Cross-modal, Data-free, Graph, Lifelong, Multi-teacher, NAS, Quantized	Image, Text	Vision-Language Models, Medical Image
This paper	Logit, Feature, Similarity	Adaptive, Adversarial, Attention, Contrastive, Cross-modal, Graph, Multi-teacher	3D/Multi-view, Image, Text, Speech, Video	Visual Recognition, Foundation Models, Transformers, LLMs, Diffusion Models, Self-supervised Learning

reviewing it from the aspects of knowledge types, algorithms, and schemes, while including comparisons between different methods. Another notable work, [12], offers an overview of model compression based on the teacher-student architecture specifically for vision tasks. [13] presents a survey on KD and proposes a new metric for comparing distillation methods based on size and accuracy, while [14] categorizes existing approaches based on the type of knowledge source used for distillation. More recently, [15] delivers a comprehensive survey, exploring knowledge optimization objectives associated with various representations, and [16] updates earlier surveys by providing an extensive review of KD methods. Additionally, they briefly investigate distillation in VLMs and discuss the challenges of distillation under limited data scenarios.

While each of these surveys provides a detailed summary of papers from different perspectives, they also have some drawbacks. First, although these surveys analyze existing approaches in each category, they fail to address recent advancements, particularly in feature-based distillation methods. Despite the significant growth of feature-based distillation, which now dominates state-of-the-art methods, many recent and impactful works are overlooked. Furthermore, adaptive distillation and contrastive distillation algorithms are rarely discussed. Second, with the recent emergence of foundation models and LLMs, these surveys have largely overlooked their potential for distilling knowledge into other models. One significant shortcomings of [11], as the most prominent survey in the field, is its lack of explanation regarding the applications of distillation in newly introduced models such as foundation models and LLMs, which have gained significant attention in recent years. Third, none of the existing surveys have investigated KD for 3D inputs like point clouds. As 3D tasks garner increasing attention in top research venues, the lack of sufficient approaches and models in comparison to the image domain highlights the importance of KD as an effective method for training the models in this area. Table I shows a summarized comparison between existing surveys and this work.

In this work, a comprehensive survey on knowledge distillation is presented. This includes reviewing existing KD methods from different perspectives: the source of distillation, distillation algorithms, distillation schemes, distillation by modalities,

and applications of KD.

The sources reviewed include logit-based, feature-based, and similarity-based distillation methods. In contrast to existing surveys, a more comprehensive classification of these methods is presented, highlighting recent advancements, particularly in feature-based distillation.

Regarding algorithms, methods in attention-based, adversarial, multi-teacher, cross-modal, graph-based, adaptive, and contrastive distillation are reviewed. Notably, adaptive and contrastive distillation are two recent categories that, despite their importance, have yet to be presented in previous works.

In terms of schemes, offline, online, and self-distillation methods are covered. Compared to previous surveys, a new section is introduced that classifies existing KD methods by modalities such as video, speech, text, multi-view, and 3D data.

In applications, a comprehensive exploration is conducted on the applications of KD in important areas including self-supervised learning, foundational models, transformers, diffusion models, visual recognition tasks, and LLMs. Finally, a quantitative comparison of prominent methods is provided, along with a discussion of current challenges and future directions.

Figure 2 shows the organization of this paper. In summary, the main contributions of this work are:

- Presenting a comprehensive survey on knowledge distillation as an essential area of research. This includes reviewing existing methods based on distillation sources, distillation algorithms, distillation schemes, modalities, and applications of distillation.
- Classifying recent distillation methods by sources, with a particular focus on feature distillation methods due to their importance and widespread application.
- Introducing two new distillation algorithms: adaptive distillation and contrastive distillation. These important categories have gained significance in distillation, especially with the advent of foundation models like CLIP, where contrastive distillation plays an increasingly crucial role.
- Reviewing the distillation methods for multi-view and 3D data. Due to the significant role of KD in 3D tasks nowadays, exploring distillation in the 3D domain is crucial,

Fig. 2. Diagram of the contents of this paper. Knowledge distillation is reviewed from different aspects, including sources, schemes, algorithms, modalities, and applications.

yet it has been overlooked in previous comprehensive surveys.

- Exploring the applications of KD in self-supervised learning, foundation models, transformers, diffusion models, and LLMs. Distillation from foundation models is a hot topic, and distillation in LLMs is particularly critical due to their large number of parameters.
- Comparing prominent distillation methods and discussing the existing challenges and future directions of KD.

II. SOURCES

The primary component of a distillation method is the source of knowledge being distilled. Various sources of knowledge can be utilized for distillation. In the initial concept of distillation [10], logits were employed as the teacher's final output. Alongside logit distillation [17]–[20], the activations [21], neurons, and features [22], [23] of intermediate layers play a crucial role in guiding the student network. Moreover, relationships and similarities among channels [24], instances [25], and classes [26] provide higher-order information that facilitates the transfer of knowledge from the teacher to the student. In this section, KD methods based on their sources of distillation are examined. Specifically, existing methods are classified into three categories: Logit-based, Feature-based, and Similarity-based as shown in Figure 3. Table II summarizes logit-based and similarity-based distillation methods and Table III presents detailed information on existing feature distillation methods. Subsequently, each type of distillation is explained in detail.

A. Logit-based Distillation

The most straightforward source for KD is logits. Logits are outputs of the last layer of the model, and the idea is to force the student model to mimic the final predictions of the teacher, which are supposed to contain informative dark knowledge [10] of the teacher. The prediction from the last layer can vary for each task. For example, logits in classification contain predictions for each class, in detection they contain predicted coordinates, and in pose estimation, they can represent heatmaps. In general, let Z_T and Z_S be logits of the last layers of the teacher and student, respectively. Then the logit-based distillation loss (L_{ld}) is defined as

$$L_{ld}(Z_T, Z_S) = \ell_{logit}(Z_T, Z_S) \tag{1}$$

where ℓ_{logit} is a metric for measuring the difference between the logits.

The first papers on KD were proposed for image classification, where soft-labels were introduced by applying softmax on logits [10], [17]. These output distributions, or soft-labels, provide information on the probability of the input belonging to each class. The distillation loss function for soft-labels is expressed as

$$p(Z_i, \tau) = \frac{exp(Z_i/\tau)}{\Sigma_i exp(Z_i/\tau)} \tag{2}$$

$$L_{ld}(p(Z_t, \tau), p(Z_s, \tau)) = KL(p(Z_t, \tau), p(Z_s, \tau)) \tag{3}$$

Fig. 3. Illustration of the sources for distillation, including logits, features, and similarities. Logits are the model's last layer output, features are intermediate outputs of the model, and similarities are relationships between features, channels, samples, etc.

where exp is softmax function, τ is the temperature factor that controls the softening of logits, and $KL(\cdot)$ is the Kullback–Leibler Divergence (KL) function. By optimizing the above objective, the student can mimic the output distribution of the teacher and make better predictions.

Although the initial concept of KD was simple and effective, many different logit-based methods have been proposed in recent years [18], [19], [27], [28]. These methods mainly aim to normalize logits before distillation [20], [29]–[32], soften or smoothen the logits [28], [33], [34], or decouple the logit-distillation loss [35]–[40]. [29] finds that the magnitude of confidence is not necessary for KD and proposes spherical KD to reduce the gap between teacher and student. [30] shows that it is not feasible to soften all the samples with a constant temperature and proposes NormKD to customize the temperature for each sample according to the characteristics of the sample's logit distribution. [31] and [32] investigate the role of projectors and temperature in the KD process respectively. [20] suggests logit standardization by setting the temperature as the weighted standard deviation of the logit and performing a plug-and-play Z-score pre-process. On the other hand, among methods that try to soften the logits, [28] perturbs the logits for distillation with some noises, and [34] softens the logits before KD and utilizes a learning simplifier with an attention module. However, most existing methods aim to decouple the distillation loss. For instance, [35]–[37] separate the distillation of target-class and non-target class distributions. [39] reshapes the logits for multi-scale logit distillation, [40] uses different weights for distilling edges and bodies of objects, and [38] decouples the KL to a weighted MSE and a cross-entropy loss that incorporates soft-labels.

In summary, logit distillation is a simple and straightforward method that can be likened to label smoothing [41] and regularization [42], [43], aiming to enhance the student network's

performance by preventing overfitting [11]. However, due to the reliance of logits on predefined classes, its application is limited to supervised learning. Moreover, mimicking the last-layer predictions does not provide insights into the intermediate layers of the teacher network. It can be challenging for the student to learn effectively solely from the final outputs of the teacher, especially when teacher and student networks have different architectures. Therefore, in addition to logit distillation, there is a necessity to distill intermediate outputs and similarities from the teacher network to facilitate effective knowledge transfer.

B. Feature-based Distillation

DNNs are proven to extract different levels of features in their layers, and these multi-level representations from the intermediate layers of a network make them a valuable source for distillation, known as feature distillation. These intermediate representations provide step-by-step information that leads to the final prediction, which can be more valuable, especially in tasks like representation learning where the output of the model is a representation rather than logits.

The concept of feature distillation was first explored by [22], where they proposed providing the student network with hints from the teacher. More specifically, they suggested minimizing the differences in features between the teacher and student to align them. Generally, let F_s and F_t be intermediate features of the teacher and student, respectively. A general feature distillation loss (L_{fd}) between teacher and student can then be defined as follows

$$L_{fd} = \ell_{feature}(\Phi_t(F_t), \Phi_s(F_s)) \tag{4}$$

where $\ell_{feature}$ is a similarity function for aligning the features, and Φ is a transformation function that matches the spatial size

TABLE II
SUMMARY OF LOGIT-BASED AND SIMILARITY-BASED DISTILLATION METHODS.

Source	Sub-category	Description	Reference
Logit	Logit Normalization	Standardizes the logits before distillation	KD [10], SphericalKD [29], NormKD [30], Miles et al. [31], TTM [32], Sun et al. [20]
	Logit Softening	Smooths the logits by adding noise	Sau et al. [28], SFKD [34], Li et al. [33]
	Decoupling	Splits the logit distillation loss into separate terms	DKL [38], BPKD [40], SDD [39], Ding et al. [37], NTCE [36]
	Revisiting Logit Distillation	Reduces the gap between teacher and student	TAKD [18], VanillaKD [19], EffDstl [27]
Similarity	Instance-level Similarity	Pairwise similarity between different data samples	MTKD [44], RKD [45], SPKD [25], IRGKD [46], Chen et al. [47], CSKD [48], Wang et al. [49], Cheng et al. [50], GIRKD [51]
	Feature/Channel-level Similarity	Pairwise similarity between features/channels	Yim et al. [52], ICKD [24], CCD [53]
	Class-level Similarity	Pairwise similarity between classes	CSKD [48], DSD [54], DistKD [55], IDD [56], AICSD [57]
	Others	Similarity between regions/pixels/neighbors/views	SKD [58], CIRKD [59], PRRD [60], BCKD [61], NRKD [62], CRLD [63]

or number of channels between the features of the teacher and the student.

Inspired by the idea of FitNet, several feature-based distillation methods have been proposed. Initially, AT [23] introduced attention transfer by minimizing the L_2 norm of the difference between the feature maps of the teacher and student. [64] focused on matching the distributions of neuron selectivity patterns between the teacher and student by minimizing the Maximum Mean Discrepancy (MMD). [21] transferred the activation boundaries formed by hidden neurons, while [65] transferred the factors of the features. [66] presented a comprehensive overhaul of feature distillation, proposing to distill the pre-activation feature maps, and [67] introduced channel-wise distillation by applying softmax to channels of features and matching the distributions between teacher and student.

Subsequently, some methods explored cross-layer feature distillation. [68] introduced Knowledge Review, where the feature maps of each layer of the teacher should match the student's feature maps in the corresponding layer and all features from the previous layers. [69] proposed a more general framework where each feature from the student should match all of the teacher's features, learning weights to determine the connection between the feature maps. This framework is a generalized version of FitNet, where the weights for corresponding features are set to one.

Recently, the focus of many papers has shifted towards the transformation function of the features, highlighting its role in feature matching beyond just matching feature sizes. [21] trains an autoencoder to transform the student's features for better alignment with the teacher's features. [70] randomly masks some pixels of the student's feature map and uses two convolution layers to transform the masked features, aligning them with the corresponding features in the teacher's model to encourage the student to produce feature maps similar to those of the teacher. [71] demonstrated that masking is not necessary and a simple Multi-Layer Perceptron (MLP) layer for transforming the student's features is sufficient.

Most recently, [72] combined the idea of diffusion [73] with KD, training a diffusion model on the teacher's feature maps and using it to denoise the student's feature maps. It considers the student's features as a noisy version of the teacher's features and employs diffusion as a transformation for the student's features. In a concurrent work, [74] used Convolutional Block Attention Mechanism (CBAM) [75] as a transformation, applying channel and spatial attention to the teacher's and student's features to highlight the most discriminative parts. It defines an MSE loss between the refined feature maps of the teacher and student.

Concurrent with the research path outlined, several other methods have been proposed. [76] rethinks raw feature alignment and decomposes the loss function of FitNets [22] into a magnitude difference term and an angular difference term. [77] matches the features of the teacher with augmented versions of the student's features, while [78] distills features in the frequency domain. [79] proposes N-to-one matching between the features of the student and teacher, respectively, and [80] suggests that distilling corresponding features results in the disregard of some layers of the teacher. Instead, they propose to fuse all the features of the teacher for distillation.

In summary, feature-based distillation methods are more generalizable as the student can access information from the intermediate layers of the teacher, providing richer knowledge. While most state-of-the-art methods are feature-based and lead to improved performance, selecting the appropriate layers for distillation, aligning corresponding features from the teacher and student, and matching feature sizes pose significant challenges in feature distillation. This is especially true in cases where the teacher and student have different architectures, a common scenario with the recent growth of foundation models. In such cases, there can be a significant decrease in performance due to the capacity gap between the teacher and student networks. Furthermore, comparing different aspects of feature-based methods is not straightforward, as their performance can vary significantly depending on factors

TABLE III
SUMMARY OF FEATURE DISTILLATION METHODS.

Method	Teacher Transformation	Student Transformation	Knowledge Type	Distillation Loss
FitNet [22] (2014)	–	1×1 Conv	Feature Representation	$\mathcal{L}_2(\cdot)$
AT [23] (2016)	Channel Aggregation	Channel Aggregation	Attention Map	$\mathcal{L}_2(\cdot)$
NST [64] (2017)	–	1×1 Conv	Neuron Selectivity Pattern	$\mathcal{L}_{MMD}(\cdot)$
Jacobian [81] (2018)	Gradient	Gradient	Jacobian of Feature	$\mathcal{L}_2(\cdot)$
FT [65] (2018)	Auto-encoder	Auto-encoder	Paraphraser	$\mathcal{L}_1(\cdot)$
AB [21] (2019)	Binarization	1×1 Conv	Activation Boundary	Marginal $\mathcal{L}_2$
Heo et al. [66] (2019)	Margin ReLU	1×1 Conv	Pre-ReLU Feature	Partial $\mathcal{L}_2$
He el al. [82] (2019)	Auto-encoder	3×3 Conv	Adapted Feature	$\mathcal{L}_1(\cdot)/\mathcal{L}_2(\cdot)$
FN [83] (2020)	L_2 Normalization	L_2 Normalization	Feature Representation	$\mathcal{L}_{CE}(\cdot)$
CWD [67] (2021)	Softmax$(\cdot)$	Softmax$(\cdot)$	Activation Map	$KL(\cdot)$
MGD [70] (2022)	–	Masking + Conv	Feature Representation	$\mathcal{L}_2(\cdot)$
MLP [71] (2023)	–	MLP	Feature Representation	$\mathcal{L}_2(\cdot)$
DiffKD [72] (2023)	–	Diffusion Process	Denoised Feature	$KL(\cdot)/\mathcal{L}_2(\cdot)$
FAKD [77] (2024)	–	Feature Augmentation	Activation Boundary	$KL(\cdot)$
LAD [76] (2024)	–	1×1 Conv	Angular Information of Feature	$\mathcal{L}_2(\cdot)$
AttnFD [74] (2024)	Channel/Spatial Attention	Channel/Spatial Attention	Refined Feature	$\mathcal{L}_2(\cdot)$

like the selection of layers, feature alignment, and feature transformation components. The sensitivity of these methods to such factors complicates their comparison and evaluation.

C. Similarity-based Distillation

In addition to logit and feature distillation, similarity distillation is another form of distillation that equips the student with the structural knowledge and relationships derived from the data learned by the teacher. Rather than relying solely on exact predictions or feature values, similarity distillation transfers a higher order of knowledge by considering the pairwise similarities between features or instances.

One of the pioneering works was [52], which introduced the Flow of Solution Procedure (FSP). This method involved computing the inner product between features from two layers in the teacher network and defining the squared L2 norm as the cost function for each pair of layers between the teacher and student networks. In a related work, [44] suggested distilling the relative dissimilarity between intermediate representations of different examples. Subsequent studies proposed distilling the 8-neighborhood similarity of logits [84] and distilling similarity of features and logits among multiple teachers [85].

In general, the similarity distillation loss (L_{SD}) is defined between a relation in the teacher and student as follows

$$L_{SD} = \ell_{similarity}(\phi(F_t^i, F_t^j), \phi(F_s^i, F_s^j)) \quad (5)$$

where, F_t and F_s represent the features/logits of the teacher and student, respectively, and F^i and F^j denote different features/logits from either two different layers or two different samples. The function ϕ calculates the similarity knowledge between the features/logits of the teacher and student, and $\ell_{similarity}$ is a correlation function that measures the similarities between the teacher and student.

Recently, similarity distillation has garnered increased attention, with numerous methods proposed, each focusing on different levels of similarity such as instance/sample-level similarity [25], [45]–[47], [49]–[51], [62], [86], feature/channel-level similarity [24], [26], [53], and class-level similarity [48], [54]–[57]. Remarkable works have emerged in instance-level methods. For example, RelationKD [45] distills the distance and anglewise correlations between features of different samples. [25] adopts a similar approach by distilling the correlation of samples within a batch through the inner-product calculation of features of each sample. [46] leverages an instance relationship graph where features are represented as nodes and relations as edges for each sample. CIRKD [59], a popular method, extends instance similarity from the batch-level by utilizing a memory bank to consider global relations between samples. [51] conducts similar work by distilling pairwise similarity relations based on stored features to reveal more complete instance relations.

In feature/channel-level similarity, IFVD [26] and ICKD [24] are two pioneering works. The former defines a prototype for each class and distills the distances of pixels from each prototype, while the latter creates a channel correlation matrix using the dot-product of a feature map with its transpose for channel-level similarity distillation. [53] compels the student to imitate the teacher by minimizing the distance between the channel correlation maps of the student and the teacher.

In class-level similarity, DistKD [55] is a popular method that significantly enhances the student's performance by distilling inter- and intra-class similarities. [54] distills the pairwise similarity of classes by multiplying the logits with its reshaped version. [56] suggests distilling position information and inter-class distances for semantic segmentation. AICSD [57] distills inter-class similarities by introducing intra-class distributions for each class and subsequently calculating pairwise similarity

of these distributions using KL for semantic segmentation.

In addition to the similarity levels mentioned, spatial similarity is also utilized for distilling pixel-to-pixel or pixel-to-region [58]–[61], [63], [87] similarities. [58] was among the first to propose pooling features into nine regions and distilling the similarity of these regions. CIRKD [59] focused on pixel-to-pixel and pixel-to-region similarities across entire images. [60] transfers the multi-scale pixel-region relation, [61] distills correlations between adjacent blocks of the teacher, [87] employs patch-level similarity for distillation, and [63] introduces distillation of the logits' similarity from local and global perspectives.

III. SCHEMES

After the knowledge source, the distillation scheme is one of the key choices in a teacher-student architecture. Based on the architecture of the teacher and its training mode, distillation methods can be categorized into three main schemes: offline distillation, online distillation, and self-distillation. In the offline scheme, which includes most of the distillation methods, the teacher is a larger pre-trained network. In the online scheme, the teacher and student are trained simultaneously, and in self-distillation, the teacher and student are the same network. Each of these schemes has its own applications: offline is preferred when a larger pre-trained teacher is available, online is most suitable when access to a pre-trained model is not available, and self-distillation plays an important role in self-supervised scenarios where labeled data is not available. Figure 4 shows the overall diagram of each scheme. In the following sections, the details of each scheme are explained.

Fig. 4. Illustration of distillation schemes. The teacher model in offline distillation is pre-trained and frozen, while in online distillation, the teacher and student are trained simultaneously. In self-distillation, the teacher and student are the same network.

A. Offline Distillation

The most straightforward approach to knowledge distillation is offline distillation, which consists of two stages. First, a large teacher model is pre-trained on a dataset. Then, during the training of the student model, the teacher's knowledge is transferred while its weights remain frozen.

Most existing distillation methods follow this offline approach, and the majority of the studies mentioned in the previous section adopt this technique. Offline distillation was first introduced in [10], and numerous methods have since been proposed to extend and improve it. However, offline distillation has some limitations. One major challenge is the memory and computational overhead required to load a large teacher model, which can be a significant issue in resource-constrained environments. Additionally, since the teacher model remains fixed, the student model's performance is inherently constrained by the teacher's capabilities. Any biases present in the teacher can also be transferred to the student, potentially affecting its generalization [29], [55].

To address these issues, incorporating adaptive distillation techniques alongside offline distillation can help improve student performance [57], [67], [88]. Moreover, in LLM distillation, where the teacher model is proprietary and only accessible via an API, offline distillation remains the only practical approach. In summary, despite its challenges, offline distillation continues to be the most widely used method in knowledge distillation due to its simplicity and effectiveness.

B. Online Distillation

Unlike offline distillation, the training of online distillation occurs in a single stage. [89] showed that online distillation offers advantages by using a group of students as a virtual teacher, eliminating the need for a large pretrained model. Early approaches, such as [90] and [91], utilized an ensemble of multi-branch logits to create an on-the-fly teacher and trained the student accordingly. In addition, [92] extended this approach by using the mean of previous predictions as an auxiliary guide to enhance learning. [93] introduced variation in branch depth, with each Hourglass network mimicking both the final heatmap and the fused heatmap of all branches. [94] proposed a framework for mutual contrastive learning, where contrastive embeddings from the same or different networks were used with the anchor during the training phase.

Recent works have focused on improving other aspects of online KD. [95] explores methods to enhance KD by reducing the sharpness and variance gap in the logit space between the teacher and student. [96] aims to achieve flatter loss minima to improve the generalization and stability of student models. [97] enhances multi-task networks by leveraging single-task networks as teachers and employing adaptive feature distillation with online task weighting. Similarly, [98] applies this approach to multi-domain tasks using multiple teachers, where each teacher model specializes in a specific domain.

C. Self-distillation

Self-distillation is a unique form of online distillation where there is no larger pre-trained teacher; instead, the teacher and student are identical networks [99]–[105]. The concept was initially introduced in [99], where a teacher is initially trained on labels, and subsequently, a new identical model is initialized from a different random seed and trained under the guidance of the former teacher. In a similar approach, [101] first trains the network with labels and then utilizes

labels and features of the model trained so far for subsequent training steps. [105] incorporates a linear combination of hard targets and predictions from the model in the previous epoch (teacher) to train the network in the current epoch (student). [100] partitions the model into several sections and transfers knowledge from deeper sections to shallower ones, while [104] employs multiple shallow classifiers at different layers and transfers knowledge from the deepest classifier to shallower ones. [103] is the first to provide a theoretical analysis of self-distillation, and [106] establishes a theoretical link between self-distillation and label smoothing.

Recently, a variety of diverse self-distillation algorithms have been proposed [107]–[117], each offering a new perspective on the problem. Some methods leverage the deeper layers of the model as teachers to distill knowledge to shallower parts of the model [116], [117]. Other approaches utilize predictions from previous epochs as teachers to supervise the model in subsequent epochs [113]–[115]. [112] and [110] employ iterative pruning and discarding of parts of the model to create a student for self-distillation. [111] and [108] integrate the concept of self-distillation with graphs, while [109] combines data augmentation, deep contrastive learning, and self-distillation. In this method, different views of the same sample are input to the model for enhanced learning.

IV. ALGORITHMS

Aside from the source and scheme, different algorithms have been proposed to effectively transfer knowledge from the teacher to the student. In this section, the most important algorithms are reviewed. These include attention-based distillation, adversarial distillation, multi-teacher distillation, cross-modal distillation, graph-based distillation, adaptive distillation, and contrastive distillation.

A. Attention-based Distillation

Attention-based distillation focuses on transferring rich information from the teacher to the student using attention mechanisms, as attention can reflect neuron selectivity. Attention maps highlight crucial regions of the features in the teacher model and enable the student to automatically and effectively focus on mimicking the important information from the teacher. Instead of exact prediction matching, this approach allows the student to concentrate on the more critical regions of the image, similar to the teacher model.

The first work exploring attention distillation was Attention Transfer (AT) [23], which proposed the aggregation of feature channels in the student and teacher models. By minimizing the loss between attention maps of the teacher and student, the attention of the teacher can be effectively transferred to the student. [54] transfers residual attention maps of features from two different layers, while [118] employs a self-attention module to adaptively aggregate context information from the student and teacher feature maps. [34] uses self-attention to simplify the learning process by softening the logits before distillation.

Several methods have focused on utilizing Class Activation Maps (CAM) [119] or Gradient-Weighted Class Activation Maps (GradCAM) [120] for attention distillation by transferring regions with more attention [121]–[123]. [124] applies channel self-attention and position self-attention on the features for distillation, [125] employs different types of attention to transfer knowledge at various levels of deep representation learning for KD, and [126] computes attention using student features as queries and teacher features as key values, implementing sparse attention values through random deactivation. These sparse attention values are then used to reweight the feature distance of each teacher-student feature pair to prevent negative transfer.

Recently, powerful attention distillation methods have been proposed [72], [74], [127]. DiffKD [72] suggests training a diffusion model on the teacher's features and using it to denoise the student's features to highlight important areas for distillation. AttnFD [74] investigates the role of attention in distillation for semantic segmentation and employs CBAM [75] attention mechanism to refine features before distillation by considering both channel and spatial attention. This refinement focuses on transferring crucial information and emphasizes important regions, particularly enhancing the student network's performance in scenarios involving small objects.

B. Adversarial Distillation

A promising direction for leveraging teacher information during distillation is to embed the process within an adversarial framework inspired by the min-max game of Generative Adversarial Networks (GANs) [128]. Three distinct approaches emerge in this context. First, generative networks can be employed to synthesize the teacher's training data when the original data is unavailable. Second, a discriminator is integrated to provide refined guidance by distinguishing between teacher and student outputs during knowledge transfer. Third, adversarial training is applied to minimize discrepancies between teacher and student models when confronted with perturbed samples, thus pushing the student to learn from more challenging examples. Collectively, these methods aim to enhance the performance and robustness of the distilled student model, and the following sections will elaborate on these three approaches in detail.

1) Adversarial Data-Free Knowledge Distillation: Adversarial distillation is applied in various ways, depending on the goal of the distillation process. One approach involves data-free distillation [129]–[141], where adversarial examples are generated in the absence of access to the teacher's original training data, often due to privacy concerns, transmission issues, or legal constraints [131]. Adversarial samples are synthesized either to mimic the original dataset or to augment it, thereby enhancing the distillation process. Methods in this category [129], [130], [132], [135], [137], [140] often leverage the pre-trained teacher network and its internal statistics as a discriminator to guide the generator. Other techniques address distribution shifts in generated samples by proposing the use of an Exponential Moving Average (EMA) of the student as a more reliable learning source [136] or framing the process as a meta-learning problem that balances knowledge acquisition and retention [134]. Self-adversarial strategies [138] have also

been explored to extend robustness in data-free settings. Lastly, [133] investigated the challenges of data-free distillation with imbalanced datasets, where the teacher model is pre-trained on skewed class distributions, introducing class-aware strategies to mitigate this imbalance.

2) Discriminator in the Loop: The use of a discriminator has proven effective in better aligning the prediction distributions between the teacher and student networks. Leveraging this intuition, one of the prevalent approaches in adversarial distillation involves incorporating a discriminator network to align the teacher and student logits [142]–[145] or intermediate feature representations [26], [58], [146]–[153]. Additionally, some methods employ a three-player framework for discrimination [154], [155], enhancing the effectiveness of the distillation process.

Adversarial distillation has also been extended beyond standard networks to the compression of GANs [156]–[159], facilitating the distillation of heavy generator networks into more efficient ones. Notably, the discrimination module is not limited to the GAN framework and has been adapted for diffusion models to reduce the number of inference steps, significantly improving their inference efficiency [160]–[163]. Furthermore, advanced techniques have been introduced to utilize the score function instead of the final outputs or logits for distillation, leading to approaches such as adversarial score distillation [164] and variational score distillation [165]. These methods employ score matching and variational approximations to achieve improved robustness and efficiency in knowledge transfer.

3) Adversarial Robust Knowledge Distillation: As previously discussed, KD transfers knowledge from a large, pre-trained teacher network to a smaller student network, yet it does not inherently guarantee robustness against adversarial attacks. Adversarial Robust Distillation (ARD) [166] addresses this limitation by combining adversarial training with distillation. In ARD, the objective is to minimize the discrepancy between the student's predictions and the teacher's predictions under adversarial perturbations, defined as follows

$$\min_{\theta} \mathbb{E}_{(X,y)\sim\mathcal{D}} \Big[\underbrace{\alpha\tau^2 D_{\mathrm{KL}}(S_\theta^\tau(X+\delta_\theta), T^\tau(X))}_{\text{Adversarially Robust Distillation loss}}$$
$$+ \underbrace{(1-\alpha)\ell(S_\theta^t(X), y)}_{\text{Classification loss}} \Big], \qquad (6)$$

where S and T denote the student and teacher networks, τ is the temperature constant, and δ_θ (computed as $\delta_\theta = \arg\max_{\|\delta\|_p < \epsilon} \ell(S_\theta^t(X), y)$) represents the adversarial perturbation.

Building upon ARD, several approaches have been developed to enhance the transfer of robust knowledge. One group refines teacher prediction reliability: RSLAD [167] replaces the classification loss in eq. (6) with a KL loss between teacher and student predictions to capitalize on robust soft labels from adversarially trained teachers, while introspective adversarial distillation (IAD) [168] and MTARD [169] further allow selective trust in teacher outputs and employ dual teachers for clean and adversarial scenarios, respectively. Another

group enhances alignment and adaptability by synchronizing gradients and adaptively deriving perturbations through IGDM [170] and AdaAD [171]; additional refinements are achieved in SmaraAD [172] and IGKD-BML [173] via Spearman Correlation, Class Activation Mapping, and attention-guided distillation to better emulate the teacher's decision-making process.

Advanced strategies tackle limitations in robustness transfer by introducing novel training schemes. DARWIN [174] incorporates intermediate adversarial samples along with a triplet-based loss to balance natural, adversarial, and intermediate distributions, while PeerAiD [175] trains a peer model alongside the student for dynamic adaptation. Furthermore, DGAD [176] partitions the dataset into standard and adversarial distillation groups with consistency regularization to manage data imbalance, and both STARSHIP [177] and TALD [178] further bolster robustness by transferring statistical attributes and sampling diverse adversarial examples via a Teacher Adversarial Local Distribution using Stein Variational Gradient Descent. These grouped techniques collectively advance the robustness and overall performance of student models in adversarial environments.

C. Multi-teacher Distillation

In contrast to typical distillation scenarios where the student is trained using a single teacher, multi-teacher algorithms aim to train the student network by amalgamating knowledge from multiple teachers. This approach has the potential to enhance the student's performance by leveraging an ensemble of teachers with diverse knowledge, thereby bolstering the student's generalization capabilities. This concept was initially proposed in [10], which utilized the average of logits from multiple teachers as softened labels. Subsequent works have explored variations on this theme: [44] employed a voting strategy, while [28] recommended perturbing the teacher's logits by adding noise, treating these perturbed outputs as distinct teachers that act as a regularizer. [179] utilized a pool of teachers, randomly selecting a teacher for distillation each time. [180] trained multiple teachers, with each focusing on a subset of the dataset, and employed a voting system to distill the knowledge of each expert teacher. [99] adopted a step-by-step strategy in which the student, at each stage, served as a teacher for the student in the subsequent step.

In recent years, a variety of multi-teacher methods have been introduced, each addressing distinct issues. Managing the knowledge aggregation adaptively poses a challenge due to the capacity gap between the student and each teacher [181]–[186]. [184] employs adaptive temperature and a diverse aggregation strategy to enhance distillation performance, while [185] separates the vanilla KD loss and assigns adaptive weights to each teacher based on the entropy of their predictions. [186] leverages data relation knowledge to dynamically allocate weights to teachers, and [182] employs a meta-network for weighting each teacher effectively.

Other methodologies have been proposed to effectively aggregate the knowledge of teachers for distillation [187]–[191]. [188] trains two distinct teachers, one for distilling

intricate features and another for transferring general decision features. [189] and [190] introduce additional branches to the student network for learning the features of teachers, while [191] suggests a progressive approach for distilling knowledge from multiple teachers. Another significant research avenue in multi-teacher distillation involves using ensemble pre-trained teachers, referred to as Knowledge Amalgamation (KA) [192]–[199]. Through KA, publicly available trained networks can be repurposed in the context of multi-teacher distillation. Some approaches involve employing multiple teachers, each trained on different datasets [174] or tasks [187], [192]. [193] initially extracts task-specific knowledge from heterogeneous teachers sharing the same sub-task and then combines this extracted knowledge to construct the student network. [199] proposes a KA framework based on uncertainty suppression, and [194] suggests training a unified classifier from pre-trained classifiers from each source along with an unlabeled set of generic data.

In summary, multi-teacher distillation furnishes the student network with a broader range of diverse and enriched information from multiple teachers, ultimately enhancing the student's generalization capabilities. Nonetheless, the effective aggregation of teachers' knowledge, determining the individual contributions of each teacher in distillation, and managing the complexity and computational costs are the primary challenges associated with multi-teacher distillation methods.

Fig. 5. General overview of a cross-modal distillation method.

D. Cross-modal Distillation

In many scenarios, some modalities have rich annotations, while others lack sufficient labels. Cross-Modality Knowledge Distillation (CMKD) addresses this challenge by leveraging a well-annotated modality to improve the learning of a less-annotated one. This technique enables the model to benefit from the additional modality during training while remaining functional, even when that modality is unavailable at inference time. In 2016, [200] introduced an innovative framework to transfer knowledge from a high-resource modality to a low-resource one, a method that has since been widely adopted in various studies. Figure 5 illustrates the overall concept of cross-modal distillation and Table IV summarizes existing methods.

The early approach for CMKD was to use two parallel streams for two modalities during training, followed by a single stream during inference to produce feature maps similar to the teacher's modality. [201], [202] applied this method

to generate depth map from RGB. Later works focused on improving the imitation of the student's output to match the teacher's and reducing the gap between the modalities [203], with [204] employing two student networks, each attempting to replicate the other's output. [205] adopted bidirectional distillation between events and frames.

The model proposed in [206] learned to generate 3D features from paired 2D inputs during the training phase and used them during inference. The structures in [207]–[210] consisted of separate encoders and decoders to extract features from LiDAR and RGB inputs, aiming to align the different features from the two modalities and distill spatial knowledge from LiDAR to RGB. [211] used the same technique, replacing LiDAR with Radar, which provided lower spatial information, alongside RGB. Instead of using voxels for alignment, [212] generated depth and semantic features from both networks and aimed to make them similar. [213] employed a shared model to extract features from Thermal and RGB inputs in the student network, while using two heavier models for the teacher. The method proposed in [214] decomposed the process into task-specific components and enabled the distillation of knowledge from optical flow effects on the right part of the network.

To distill knowledge between text and image modalities, [215] extracted features separately and trained them on aligned instances, using soft alignment to supervise the student network. [216] applied masking techniques and used text features to guide the network in predicting masked images. [217] masked audio and video frames, attempting to reconstruct them to achieve a good representation for both modalities and align the two features effectively.

TABLE IV
SUMMARY OF CROSS-MODAL DISTILLATION METHODS.

Method	Train Modality	Inference Modality
Gupta et al. [200] (2016)	RGB	Depth/Optical Flow
Garcia et al. [202] (2019)	RGB + Depth	RGB
EvDistill [205] (2021)	RGB + Event	Event
Hong et al. [207] (2022) DistillBEV [209] (2023)	RGB + LiDAR	RGB
UniDistill [208] (2023)	RGB + LiDAR	LiDAR
Lee et al. [214] (2023)	RGB + Optical Flow	RGB
Radocc [212] (2024)	Multi-view + LiDAR	Mult-view
CRKD [211] (2024)	Multi-view + LiDAR	Multi-view + Radar
Andonian et al. [215] (2022) CM-MaskSD [216] (2024)	RGB + Text	RGB + Text
XKD [217] (2024)	Video	RGB

E. Graph-based Distillation

Graph-based knowledge distillation extends traditional KD methods by incorporating higher-order relational information between features or instances. Unlike approaches that primarily rely on pairwise similarity, graph-based methods utilize graph structures as a generic framework to capture intra-dependent relationships. In such methods, features or instances are represented as vertices, while the dependencies between them are modeled through edge weights. These graph structures encapsulate both local and global relationships, enabling

the transfer of richer and more structured knowledge from teacher models to student models [25], [45]–[47], [85], [86], [218]–[223]. Furthermore, graph structures can be effectively employed to model the flow of information during distillation, particularly in scenarios involving multiple teacher models, by regulating the knowledge transfer process [224], [225].

In recent years, the advent of Graph Neural Networks (GNNs) has garnered significant attention, particularly in the fields of data mining and knowledge graph modeling. GNN distillation is primarily utilized to achieve two main objectives [226]: enhancing the performance of the original teacher model (performance improvement) [227]–[235] or creating a lightweight version of the GNN (compression). The latter involves distilling a larger and more complex GNN into a compact and efficient GNN [236]–[239] or even an MLP [240]–[243], making the distilled model suitable for real-time applications. To accomplish these goals, the information transferred between the teacher and the student typically includes logits [227], [228], [231], [232], [236], [238]–[241], [243], feature representations [227], [230], [232], [234], [238], and the graph structure itself [228]–[230], [235], [237], [242], [243], which encapsulates the connectivity between the elements of the graph. These components collectively enable an effective transfer of structured knowledge, ensuring that the distilled model retains essential characteristics while meeting its respective performance or efficiency objectives.

F. Adaptive Distillation

Adaptive distillation has recently garnered increased attention. By dynamically adjusting the parameters of the distillation process instead of maintaining a constant setup, the effectiveness of knowledge transfer can be enhanced. Table V summarizes adaptive distillation methods in different categories.

Adaptive distillation can manifest in various levels and forms, such as adaptive loss definition [49], [244], [245], adaptive loss weighting [57], [88], [246], adaptive teacher pruning [18], [219], [247]–[249], and dynamically weighting different aspects of distillation based on sample importance [250]–[255].

One primary concern is that the teacher network may produce incorrect outputs that should not be directly transferred to the student. To address this issue, some approaches have introduced adaptive losses. [244] suggests an adaptive Cross Entropy loss that replaces incorrect probability maps with ground truth labels. [49] proposes a calibrated mask to prevent the teacher model's erroneous representations from interfering with the student model's training, while [245] integrates the positive prediction distribution of two teacher networks based on their correctness and cross-entropy magnitudes to provide a more accurate output distribution for guiding the student network.

Moreover, certain studies have delved into the effects of altering the distillation loss throughout the distillation process. [88] was among the first to explore strategies where reducing the influence of the teacher network on the student can enhance performance. [246] indicates that the impact of the distillation

loss should diminish as training progresses, with the student gradually assuming control of the training process towards the end. [57] introduces a weight decaying process to merge similarity distillation with vanilla distillation.

Given the capacity disparity between teacher and student models, some works aim to narrow this gap. A notable example is TAKD [18], which employs a teaching assistant network as an intermediary teacher to facilitate better knowledge comprehension by the student model. Similarly, [219] utilizes an auxiliary teacher model before distillation, while [248] and [249] advocate for channel and feature pruning, respectively, before distillation. [247] implements filter pruning to reduce channels and applies a curriculum learning strategy to distill layers from easy to challenging levels.

Lastly, certain methodologies adaptively conduct distillation based on the significance of a region or sample. [251] extracts detailed context-specific information from each training sample, [253] adaptively identifies confusing samples, and [254] adjusts the temperature parameter based on each sample's energy level. [250] incorporates adaptive weights for distilling each mask in feature distillation, while [255] employs ground truth information to mask crucial regions for logit distillation. [252] calculates channel divergences between the teacher and student networks, converting them into distillation weights, and [256] assigns greater weight to losses in layers where training discrepancies between the teacher and student models are more pronounced during distillation.

TABLE V
SUMMARY OF ADAPTIVE DISTILLATION METHODS.

Method	Reference
Adaptive Loss Definition	Park et al. [244], Wang et al. [49], CAG-DAKD [245]
Adaptive Loss Scaling	Cho et al. [88], Zhou et al. [246], AICSD [57]
Adaptive Teacher Pruning	TAKD [18], InDistill [247], Passalis et al. [219], PCD [248], AFMPM [249]
Adaptive Sample Importance	MasKD [250], APD [251], HWD [252], CS-KD [253], Energy KD [254], Gou et al. [255]

G. Contrastive Distillation

Contrastive learning seeks to learn representations by contrasting similar (positive) and dissimilar (negative) samples in the feature space [257]. In KD, most research employs feature-based contrastive learning approaches, where anchor, positive, and negative sets are defined using features extracted by teacher and student models as shown in Figure 6. The contrastive distillation loss can be formulated as

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(F_i^s, F_i^t)/\tau)}{\sum_{j \in \mathcal{N}} \exp(\text{sim}(F_i^s, F_j)/\tau)}, \quad (7)$$

where anchors (F_i^s) are typically features from the student's embeddings, positives (F_i^t) are corresponding features from the teacher for the same input, and negatives (F_j) are unrelated features from other data points. These methods employ

contrastive loss to align the student's feature space with the teacher's, enabling efficient knowledge transfer. Other studies [258], [259] innovate by introducing new contrastive loss functions that address challenges such as noisy negatives and capacity mismatches between teacher and student models, enhancing the robustness and effectiveness of the distillation process.

Fig. 6. General overview of a contrastive distillation method. Embeddings of the teacher and student are considered as the positive set and anchor, respectively. The negative set can consist of embeddings from the teacher, student, or input data.

Contrastive distillation has gained significant attention for its ability to enhance knowledge transfer through contrastive learning techniques. Tian et al. [260] pioneered this approach by employing a contrastive loss to align the features of the teacher and student models, Expanding upon this foundation, WCoRD [261] introduced a method that incorporates both primal and dual forms of the Wasserstein distance. While the dual form ensures global knowledge transfer by maximizing mutual information, the primal form focuses on local feature alignment within mini-batches, providing a balanced approach to feature alignment and distillation.

In the field of semantic segmentation, [259] addresses the challenge of dense pixel-level representations by leveraging contrastive loss to align the teacher and student feature maps. Similarly, [262] focuses on the domain of single-image super-resolution. Their approach utilizes contrastive loss to distill statistical information from intermediate teacher features into a lightweight student network, enabling high-resolution image reconstruction with minimal computational resources.

The problem of distilling large language-image pretraining models, such as CLIP [263], is tackled by [264], which develops a novel framework incorporating image feature alignment and educational attention modules. This method effectively aligns multi-modal features through contrastive loss. To address challenges in sentence embedding, DistilCSE [258] proposes a method which compresses large contrastive sentence embedding models.

In the domain of medical imaging, CRCKD [265] combines class-guided contrastive distillation and categorical relation preserving techniques to enhance intra-class similarity and inter-class divergence. The use of class centroids and relational graphs ensures effective knowledge transfer, even in imbalanced datasets.

In the area of weakly supervised visual grounding, [266] uses pseudo labels generated by object detectors. By applying contrastive loss for region-phrase matching, their method eliminates the need for object detection during inference, simplifying the process. [267] tackles the problem of relation-based knowledge transfer with CRCD. By maximizing mutual information between anchor-teacher and anchor-student relation distributions, their framework aligns inter-sample relations through a relation contrastive loss.

In conclusion, the incorporation of contrastive learning into distillation demonstrates exceptional capabilities in feature alignment, relational knowledge preservation, and the integration of local and global information. Proficiency of contrastive distillation in addressing challenges such as multi-modal alignment, dense feature mapping, and efficient model compression underscores its versatility. Moreover, the inherent flexibility of contrastive learning-based distillation allows it to effectively address critical issues, including imbalanced datasets, noisy labels, and computational constraints.

V. MODALITIES

Although most distillation methods have been proposed for computer vision tasks and work with images, KD has been effectively applied to other modalities, such as 3D/multi-view data, text, speech, and video. This section provides a review of distillation methods for each modality, categorizing them based on their respective tasks.

A. 3D Input

Knowledge distillation techniques contribute significantly to enhancing various 3D-related tasks, including object detection, semantic segmentation, shape generation, and shape classification. This section explores the application of KD in 3D data, highlighting innovative approaches and categorizing contributions based on task domains. Figure 7 illustrates the key tasks, domains, and types of 3D data where KD techniques have been successfully applied. Table VI summarizes distillation methods for each task based on the source of distillation.

Fig. 7. Overview of 3D domains and tasks.

1) 3D Object Detection: 3D object detection involves identifying and localizing objects within three-dimensional spaces using data such as point clouds, voxel grids, and images from multiple angles. KD enhances this process by

transferring knowledge from more complex models, leveraging various approaches and modalities, to simpler ones. This approach improves accuracy and computational efficiency in tasks including autonomous driving and multi-camera detection. Recent advancements in KD-based 3D object detection have focused on three major areas: LiDAR-based, camera-only (single and multi-camera), and cross-modal, each addressing unique challenges in 3D perception.

LiDAR: Precise spatial and depth sensing are essential for autonomous systems, , yet challenges such as sparsity and occlusion often degrade performance in 3D object detection. To solve these, several KD frameworks have been proposed [268]–[276]. X-Ray Distillation [274] trains a teacher model on object-complete frames that are derived from aggregated LiDAR scans, to transfer knowledge to a student model that improves its ability to handle sparse data and occlusions. Similarly, RadarDistill [275] aligns radar and LiDAR features to handle sparsity and noise. Furthermore, RDD [271] reduces the mismatch between the teacher and student models by aligning feature representations and refining outputs.

In addition, PointDistiller [272] focuses on local geometric structures in point clouds, using dynamic graph convolution and reweighted learning to focus on important points and voxels. itKD [273] applies feature reduction and mutual refinement to transfer coarse and fine-grained geometric features. Additionally, SRKD [276] bridges domain gaps caused by weather variations by aligning instance features through density and shape similarity and ensuring consistent predictions between the teacher and student models. LiDAR Distillation [268] also addresses domain gaps caused by varying LiDAR beam densities by performing distillation from a high-beam LiDAR teacher to a low-beam student.

Moreover, another study [269] introduces two novel techniques: pivotal position logit KD, which focuses on key areas for enhanced distillation, and teacher-guided initialization, which transfers the teacher's feature extraction capabilities to the student through weight inheritance. However, intrinsic challenges such as sparsity, randomness, and varying density limit the effectiveness of normal distillation algorithms. To overcome these limitations, PVD [270] leverages the advantages of both point-level and voxel-level knowledge to address these challenges in LiDAR object detection tasks.

Camera-only (single or multi-camera): Achieving accurate 3D object detection using only cameras is challenging due to the absence of depth information. To address this, FD3D [277] uses selective masked generative distillation and query-based focal distillation to enhance object-specific learning in perspective and Bird's-Eye View (BEV) spaces. Similarly, X3KD [210] integrates cross-task, cross-modal, and cross-stage distillation with techniques such as LiDAR feature alignment and adversarial training for dense supervision.

Cross-modal: Transferring spatial and depth knowledge from LiDAR-based models to cost-effective camera-radar or camera-only alternatives enhances multi-sensor systems, improving detection performance while resolving sensor inconsistencies. CRKD [211] transfers knowledge from LiDAR-camera teacher to a camera-radar student in the BEV space. Similarly, DistillBEV [209] transfers 3D geometric knowl-

edge from a LiDAR-based teacher to a multi-camera BEV student using region decomposition, adaptive scaling, spatial attention, and multi-scale distillation. In addition, CMKD [207] transfers spatial knowledge from LiDAR-based teachers to monocular students, tackling depth estimation challenges using BEV features and teacher predictions. In addition UniDistill [208] proposes a universal KD framework which supports multiple modality pairs (LiDAR-to-camera, camera-to-LiDAR, fusion-to-camera, and fusion-to-LiDAR). The model employed BEV as a shared representation to align teacher and student detectors across modalities.

2) 3D Shape Classification: Point cloud classification is another critical task where the teacher-student framework can be applied. FAD [278] introduces a novel loss function that combines logit distillation and feature distillation. JGEKD [281] proposes a loss function based on joint graph entropy to address the challenges of non-independent and identically distributed 3D point cloud data in classification tasks. To improve the efficiency, Zhang et al. in [280] introduced an offline distillation framework that incorporates a negative-weight self-distillation approach. [279] employs a self-distillation framework for incomplete point cloud classification.

3) 3D Semantic Segmentation: KD enhances the accuracy and robustness of point cloud segmentation by effectively transferring knowledge from a high-capacity teacher model to a lightweight student model [124], [206], [282]–[291].

In order to address challenges such as sparsity, randomness and varying density in segmentation, [287] proposes point-to-voxel KD which distills the probabilistic outputs of the teacher at both the point level (fine-grained details) and voxel level (coarse structural information). Qiu et al. [282] proposes a multi-to-single KD framework that fuses multi-scan information for hard classes and employs a multilevel distillation strategy, including feature, logit, and instance-aware similarity distillation.

Limited labeled data is another major challenge in the semantic segmentation of large-scale point clouds. Despite the task of annotating pointwise labels for such datasets being expensive and time-consuming, existing methods often rely on fully supervised approaches that require dense annotations. PSD [289] addresses this issue by introducing perturbed branches and leveraging graph-based consistency in a self-distillation manner. Seal [286] is a self-supervised learning framework that distills knowledge from off-the-shelf vision foundation models for point cloud segmentation. PartDistill [291] proposes a method to distill knowledge from a vision-language model to a 3D segmentation network.

4) 3D Domain Adaptation: Unsupervised Domain Adaptation (UDA) struggles to perform effectively when the target domain differs from the labeled source domain, due to factors such as noise, occlusions, or missing elements. [309] proposes a self-distillation UDA technique to generate discriminative representations for the target domain and utilizes GNN-based online pseudo-label refinement. [310] introduces a cross-modal feature fusion in UDA for semantic segmentation. [311] proposes a self-ensembling network for domain adaptation in 3D point clouds, which leverages a semi-supervised learning

TABLE VI
SUMMARY OF 3D DISTILLATION METHODS BASED ON THEIR TASKS AND SOURCES OF DISTILLATION.

Task	Feature-based	Similarity-based	Logit-based
Object Detection	FD3D [277], X3KD [210], Gambashidze et al. [274], PointDistiller [272], itKD [273], CRKD [211], DistillBEV [209], RDD [271], RadarDistill [275], SRKD [276]	FD3D [277], X3KD [210], Gambashidze et al. [274], itKD [273], CRKD [211]	X3KD [210], Gambashidze et al. [274], PointDistiller [272], CRKD [211], DistillBEV [209], RDD [271]
Classification	FAD [278], HSD [279]	–	FAD [278], JGEKD [280], PointViG-Distil [281]
Segmentation	Liu et al. [206], Qiu et al. [282], CMDFusion [283], Jiang et al. [284], Smaller3d [285], Seal [286]	PVD [287], Qiu et al. [282]	Genova et al. [288], PSD [289], PVD [287], Qiu et al. [282], LGKD [290], Smaller3d [285], PartDistill [291]
Depth Estimation	MVP-Net [292], LiRCDepth [293]	LiRCDepth [293]	MVP-Net [292], KD-MonoRec [294], LiRCDepth [293]
Representation	PPKT [295], Fu et al. [296], SLidR [297], PointCMT [298], RECON [299], MVNet [300], LiDAR2Map [301], C2P [302], HVDistill [303], Yao et al. [304], Text4Point [305], Diff3F [306]	–	Yu et al. [307], LiDAR2Map [301], I2P-MAE [308]
Domain Adaptation	Cardace et al. [309], Wu et al. [310], SEN [311]	–	SEN [311]
Recognition	DistilVPR [312], PointMCD [313]	–	–
Completion	SCPNet [314], RaPD [315], Lin et al. [316], VPNet [317], Huang et al. [318]	ADNet [319]	Hwang et al. [320], MonDi [321], Zhou et al. [322], Zhang et al. [323], HASSC [324], Huang et al. [318]
Registration	Jiang et al. [325]	–	DiReg [326]
Other 3D Tasks	DTC123 [327], CSD [328], Van et al. [329]	Van et al. [329]	PCDNet [330]

framework for effective knowledge transfer from a labeled source domain to an unlabeled target domain.

5) 3D Depth Estimation: 3D depth estimation is crucial for accurately understanding the geometry of a scene, enabling perception systems to reconstruct spatial relationships and object structures. MVP-Net [292] enhances reconstruction by using depth estimation in a multi-view, cross-modal distillation approach for better point cloud upsampling. KD-MonoRec [294] leverages monocular RGB images as input and employs KD along with point cloud optimization to improve depth estimation. LiRCDepth [293] utilizes RGB images and radar point clouds and proposes a lightweight depth estimation model via KD.

6) 3D Representation Learning: Representation learning focuses on extracting meaningful representations from data. Collecting a large and high-quality annotated dataset in 3D is a costly and time-consuming process, and pre-training 3D models requires access to large-scale 3D datasets, which is not a trivial task. However, some works, such as DCGLR [296], have effectively utilized knowledge distillation (KD) with only 3D data to address this challenge. However, given that numerous large-scale datasets and powerful foundation models are available in the 2D domain, several methods [295], [297], [298], [300], [301], [303], [304], [306]–[308] have leveraged these capabilities to learn rich representations for 3D point clouds.

[295] introduces contrastive learning to transfer 2D semantic knowledge to 3D networks. SLidR [297] introduces a self-supervised 2D-to-3D representation distillation framework that uses superpixel-driven contrastive loss to align image and LiDAR features, enabling effective pre-training of 3D perception models for autonomous driving. [299] proposes a framework that can be used in either a single-modal or cross-modal setting.

3D representations can be learned with the help of text [305] or through sequences of point clouds, as demonstrated in [302], which develops a self-supervised method for learning 4D point cloud representations using KD.

7) 3D Recognition: Visual place recognition is an important task in computer vision and robotics, aiming to identify previously visited locations based on visual input such as camera image and LiDAR point clouds. PointMCD [313] and DistilVPR [312] are two innovative cross-modal KD frameworks in this field, with PointMCD employing a multi-view framework, as discussed further in Section V-B.

8) 3D Completion: Point cloud completion refers to the process of restoring missing geometric details. KD helps transfer learned priors from unpaired data, reducing the need for large annotated datasets and improving completion performance in data-scarce scenarios.

Shape Completion: Shape completion aims to reconstruct a complete point cloud from occluded input point cloud.

RaPD [315] is the first semi-supervised point cloud completion method that utilizes prior distillation. [316] uses loss distillation and [322] propose a hierarchical self-distillation point cloud completion method.

Scene Completion: LiDAR sensors often capture incomplete point clouds due to occlusions from objects or their surroundings, making large-scale 3D scene interpretation challenging. KD can enhance their performance as mentioned in [314], [317], [323], [324].

Depth Completion: Depth completion is the process of estimating a dense depth map from sparse or incomplete depth measurements, such as those obtained from LiDAR sensors. KD can aid in predicting missing depth information [318]–[321].

9) 3D Registration: KD can enhance point cloud registration by transferring knowledge from a large, complex model to a smaller, efficient one while preserving performance [325], [326].

10) Other 3D Tasks: KD is applied to various other 3D tasks with different types of distillation. This includes generation [327], stylization [328], sampling [330] and accordance detection, as in [329], which utilizes all types of KD.

B. Multi-view Input

In recent years, multi-view learning has emerged as a powerful approach for addressing complex tasks, such as 3D object detection and reconstruction. While leveraging information from multiple perspectives enables models to achieve more accurate and robust results, the diversity of data modalities and the challenge of transferring knowledge across them present significant challenges. To address these, several studies have explored novel KD techniques that allow models to share and refine knowledge across different views or modalities. This section reviews recent advancements in multi-view distillation methods, categorizing them by task and the modality. The existing works can broadly be divided into two main groups: 3D object detection and 3D shape recognition, with other tasks represented by individual studies.

1) 3D Object Detection: 3D object detection has been a focal point in multi-view learning, with numerous studies proposing novel distillation techniques to enhance performance [209], [210], [333]–[335], [340], [341]. To this end, SimDistill [341] introduces a LiDAR-camera fusion-based teacher and a simulated fusion-based student for multi-modal learning. By maintaining identical architectures and incorporating a geometry compensation module, the student learns to generate multi-modal features solely from multi-view images.

[210] proposes a comprehensive framework for multi-camera 3D object detection by leveraging cross-task and cross-modal distillation. Cross-task distillation from an instance segmentation teacher avoids ambiguous error propagation, while cross-modal feature distillation and adversarial training refine 3D representations using a LiDAR-based teacher. Cross-modal output distillation further enhances detection accuracy.

[209] proposes a KD framework for aligning a BEV-based student's features with a LiDAR-based teacher's, addressing the depth and geometry inference issue compared to LiDAR-based methods.

Cross-modal KD often suffers from feature distribution mismatches. FSD scheme [340] eliminates the need for pre-trained teachers and complex strategies.

BEV-LGKD [333] proposes a LiDAR-guided KD framework for multi-view BEV 3D object detection. By transforming LiDAR points into BEV space and generating foreground masks, the method guides RGB-based BEV models without requiring LiDAR at inference. Depth distillation is also incorporated to improve depth estimation, enhancing BEV perception performance.

BEVDistill [334] introduces a cross-modal BEV distillation framework for multi-view 3D object detection by unifying image and LiDAR features in BEV space and adaptively transfers knowledge across non-homogeneous representations in a teacher-student paradigm.

STXD [335] proposes a structural and temporal cross-modal distillation framework for multi-view 3D object detection. It reduces redundancy in the student's feature components by regularizing cross-correlation while maximizing cross-modal similarities. Additionally, it encodes temporal relations across frames using similarity maps and employs response distillation to enhance output-level knowledge transfer.

2) 3D Shape Recognition: 3D shape recognition is another critical area where multi-view distillation has been applied to bridge the gap between 2D and 3D representations [313], [342]. PointMCD [313] bridges 2D visual and 3D geometric domains through a multi-view cross-modal distillation framework, by distilling knowledge from the teacher to a point encoder student. Group Multi-View Transformer (GMViT) [342] addresses the limitations of view-based 3D shape recognition methods, which often struggle with large model sizes that are unsuitable for memory-limited devices. To overcome this, GMViT introduces a large high-performing model designed to enhance the capabilities of smaller student models

3) Other Tasks: In addition to 3D object detection and 3D shape recognition, multi-view distillation techniques have been applied to a range of other tasks, including multi-view image classification [332], vision-based robotic manipulation [336], multi-view BEV detection [337], multi-view stereo depth reconstruction [331], semantic segmentation [338], human action recognition [339], and medical applications [343]. These studies are summarized as follows:

MTS-Net [332] integrates KD with multi-view learning, redefining the roles of the teacher and student models. This end-to-end approach optimizes both view classification and knowledge transfer, with extensions like MTSCNN refining multi-view feature learning for image recognition.

In robotic manipulation, [336] improves vision-based reinforcement learning by transferring knowledge from a teacher policy trained with multiple camera viewpoints to a student policy using a single viewpoint.

[337] uses a spatio-temporal distillation and BEV response distillation by aligning the student's outputs with those of the teacher, addressing the computational complexity of multi-view BEV detection methods.

Supervised multi-view stereo methods face challenges due to the scarcity of ground-truth depth data. The self-supervised KD-MVS framework [331] uses a teacher trained with pho-

TABLE VII
SUMMARY OF KNOWLEDGE DISTILLATION IN MULTI-VIEW TASKS.

Method	Task	Teacher Modality	Student Modality	Short Description
KD-MVS [331] (2022)	Multi-view Stereo Depth Reconstruction	Multi-view Images	Multi-view Images	Self-supervised MVS distillation for depth reconstruction.
MTS-Net [332] (2022)	Multi-view Image Classification	Multi-view Images	Multi-view Images	Multi-view learning and distillation for image classification.
BEV-LGKD [333] (2022)	Multi-view BEV 3D Object Detection	LiDAR	Multi-view RGB Images	LiDAR-guided distillation with foreground masks and depth distillation for BEV perception.
BEVDistill [334] (2022)	Multi-view 3D Object Detection	LiDAR	Multi-view Images	Cross-modal BEV distillation for unifying image and LiDAR features.
STXD [335] (2023)	Multi-view 3D Object Detection	LiDAR	Multi-view Images	Structural and temporal distillation with cross-correlation regularization and response distillation.
PointMCD [313] (2023)	3D Shape Recognition	Multi-view Images	3D Point Cloud	Multi-view cross-modal distillation for 3D shape recognition.
Acar et al. [336] (2023)	Vision-based Robotic Manipulation	Multi-view Images	Single-camera	Distillation from multi-camera to single-camera for robust manipulation.
Zhang et al. [337] (2023)	Multi-view BEV Detection	Multi-view BEV Images	Multi-view BEV Images	Structured distillation for efficient BEV detection.
MVKD [338] (2023)	Semantic Segmentation	Multi-view Images	Single-view Images	Multi-view distillation with co-tuning and feature/output distillation losses.
MKDT [339] (2023)	Human Action Recognition	Multi-view Videos	Single-view Videos	Multi-view distillation with hierarchical vision transformer for incomplete data.
3X3KD [210] (2023)	Multi-camera 3D Object Detection	LiDAR	Multi-view Images	Cross-modal distillation for multi-camera 3D object detection.
DistillBEV [209] (2023)	3D Object Detection for Autonomous Driving	LiDAR	Multi-view BEV Images	Distillation from LiDAR to BEV for enhanced 3D perception.
FSD-BEV [340] (2024)	3D Object Detection for Autonomous Driving	Self-distillation	Self-distillation	Foreground self-distillation for single-model distillation.
SimDistill [341] (2024)	Multi-modal 3D Object Detection	LiDAR-camera Fusion	Multi-view Images	Distillation for 3D object detection with modality gap compensation.
GMViT [342] (2024)	3D Shape Recognition	Multi-view Images	Multi-view Images	Knowledge distillation with GMViT for efficient 3D shape analysis.
MT-MV-KDF [343] (2024)	Myocardial Infarction Detection and Localization	Multi-view ECG	Single-view ECG	Multi-task, multi-view distillation with CNN and attention modules.

tometric and featuremetric consistency to probabilistically transfer knowledge to a student.

MVKD [338] introduces a multi-view KD framework for efficient semantic segmentation. The framework aggregates knowledge from multiple teacher models and transfers it to a student model. To ensure consistency among the multi-view knowledge, MVKD employs a multi-view co-tuning strategy. Additionally, it proposes multi-view feature distillation loss and multi-view output distillation loss to effectively transfer knowledge from multiple teachers to the student.

MKDT [339] introduces a multi-view KD transformer framework for human action recognition, addressing the challenge of incomplete multi-view data. The framework consists of a teacher network and a student network, both utilizing a hierarchical vision transformer with shifted windows to effectively capture spatio-temporal information.

MT-MV-KDF [343] introduces a novel multi-task multi-view KD framework for myocardial infarction detection and localization. Multi-view learning extracts ECG feature representations from different views, while multi-task learning captures similarities and differences across related tasks.

In summary, multi-view distillation techniques have shown significant potential across a broad range of tasks, from 3D object detection and shape recognition to specialized applications such as robotic manipulation and medical diagnostics. The variety of approaches, as highlighted in Table VII, emphasizes the adaptability of these methods to various modalities and challenges.

C. Text Input

Knowledge distillation has been widely applied across various NLP tasks. This technique is particularly valuable in NLP due to the large scale of state-of-the-art models, namely BERT and GPT. While these models are highly accurate, they can be impractical for deployment in resource-constrained environments. By distilling the rich representations learned by these large models into lightweight versions, KD helps maintain high accuracy in tasks such as neural machine translation, question answering, text generation, event detection, document retrieval, text recognition, named entity recognition, text summarization, natural language understanding, sentiment analysis, and text classification.

Neural Machine Translation (NMT): While NMT has significantly outperformed traditional statistical methods, large models like transformers are computationally intensive. KD addresses this challenge by distilling the teacher's output to the student, achieving comparable translation quality with reduced computational costs. This enables the development of more efficient and scalable NMT solutions [344]–[348]

Question Answering (QA): In QA, KD transfers both answer predictions and intermediate representations, such as attention distributions, from a large teacher model to a smaller student model. This process helps maintain high accuracy

while minimizing computational costs, making QA systems more efficient and suitable for real-world applications [349]–[353].

Text Generation: In text generation, KD enables the transfer of a teacher's generative capabilities to a smaller student model, training it to mimic the vocabulary distributions and language patterns of the teacher. [354] focuses on leveraging BERT's contextual understanding for efficient text generation, while [355] enhances fluency and diversity by combining distillation with GANs.

Event Detection: KD is applied in event detection to distill knowledge from larger, more complex models to smaller, more efficient ones. [356] improves event detection across languages by distilling knowledge from high-resource to low-resource models, while [357] uses distillation to retain learned knowledge while adapting to new events. Meanwhile, [143] employs adversarial imitation to enhance detection accuracy.

Document Retrieval: In document retrieval, KD transfers the ranking and matching capabilities of large teacher models to smaller student models, typically replicating relevance scores and interaction features [358], [359].

Text Recognition: KD in text recognition facilitates the transfer of both visual and contextual features from a teacher to a student, allowing the student to maintain high recognition accuracy with reduced computational cost [360], [361].

Named Entity Recognition (NER): In NER, KD helps a smaller student model learn from the teacher's contextual understanding and classification abilities [362], [363].

Text Summarization: KD in text summarization transfers summarization capabilities from teacher to student, training the student to replicate output sequences and probability distributions [364]–[366].

Natural Language Understanding (NLU): KD in NLU transfers the rich linguistic understanding of the teacher models to the student. In NLU, the student learns to replicate the teacher's output probabilities and internal representations across various tasks, ensuring high performance while reducing computational costs [367]–[369].

Sentiment Analysis: In sentiment analysis, KD distills the sentiment classification capabilities of the teacher to the student, training it to mimic the teacher's probability distributions and decision patterns [370], [371].

Text Classification: In text classification, KD transfers classification capabilities from the teacher to the student, enabling the student to mimic the teacher's feature representations and output probabilities [372], [373].

D. Speech Input

Deep neural acoustic models have achieved remarkable success in speech recognition tasks; however, their complexity poses challenges for real-time deployment on devices with limited computational resources. As the demand for efficient and rapid processing increases on embedded platforms, conventional large-scale models often increases on embedded platforms. To this end, KD has garnered significant attention in various speech-related tasks, including speech recognition, speech enhancement, speaker recognition and verification, speech translation, speech synthesis, speech separation, spoken language identification and understanding, deepfake speech and spoofing detection, audio classification and tagging, spoken question answering and conversational AI, and audio captioning and retrieval.

Speech Recognition (ASR): KD is important in enhancing ASR models by transferring expertise from high-precision teacher models to student models, supporting rapid inference with minimal loss in accuracy. This also improves domain adaptation, making ASR effective in poor acoustic environments or in expert domains, including medical transcription [374]–[385].

Speech Enhancement: KD in speech enhancement enables denoising through high-fidelity speech representations and model compression [386]–[390].

Speaker Recognition and Verification: In speaker recognition and verification, KD helps reduce model complexity while preserving speaker identity features, allowing for fast and secure authentication. It also enhances robustness against adversarial attacks and cross-domain variations, improving performance in noisy or multilingual environments [391]–[397].

Speech Translation: KD improves the efficiency of multilingual translation models by enabling knowledge sharing between high-resource and low-resource languages [398]–[401]. This enhances translation fluency while maintaining low latency for real-time applications.

Speech Synthesis (Text-to-Speech): In speech synthesis, KD reduces the complexity of deep generative models while retaining natural-sounding speech output [402]–[404]. It also helps distill expressive features, making synthetic voices more human-like with lower computation costs.

Speech Separation: KD compresses complex separation models into smaller networks while preserving speaker distinction [405], [406]. It enables real-time speaker and cocktail-party effect reduction on low-power devices.

Spoken Language Identification and Understanding: In this task, KD enables language adaption in low-resource languages through the transfer of information between high-resource multilingual teacher models [407]–[413]. As a result, such mechanism empowers lighter, accelerated models to achieve high performance in language identification and speech dialogue comprehension.

Deepfake Speech and Spoofing Detection: With growing concerns about synthetic voice fraud, the need for reliable approaches to deepfake detection has increased. KD strengthens anti-spoofing by transferring knowledge between DNNs trained on a range of deepfake speech datasets and lightweight detection architectures [414]–[419].

Audio Classification and Tagging: KD strengthens feature learning relevant to sound event detection, enhancing model efficiency and allowing deployment on embedded platforms such as IoT and edge AI platforms [420]–[425].

Spoken Question Answering and Conversational AI: KD enables efficient question-answer through the distillation of contextual information in deep, lengthy transformer models, mapping them onto efficient, quick-answer-producing models [380], [426]–[428].

Audio Captioning and Retrieval: In this task, KD enables deep audio-text embedding models to be compressed while maintaining generalization, making them suitable for efficient multimedia search architectures [429], [430].

TABLE VIII
SUMMARY OF VIDEO KNOWLEDGE DISTILLATION METHODS BASED ON THEIR SOURCES OF DISTILLATION.

Category	Method
Feature-based	Wang et al. [431], Liu et al. [432], V2I-DETR [433], MaskAgain [434]
Similarity-based	RSKD [435], Bhardwaj et al. [436], OOKD [437], DL-DKD [438]
Logit-based	Camarena et al. [439], Wu et al. [440], PKD [441], MobileVOS [442], V2I-DETR [433], DTO [443], MVD [444], RankDVQA-mini [445], VideoAdviser [446], Afouras et al. [447], Perez et al. [448]

E. Video Input

Knowledge distillation has been widely applied in image and language tasks, but its potential in video-based applications is gaining increasing attention. Recent studies have explored KD for video action recognition [431], [432], [439]–[441], video classification [435], [436], video object segmentation [433], [442], video instance segmentation [437], Partially Relevant Video Retrieval (PRVR) [438], trajectory forecasting [443], and representation learning through masked video modeling [434], [444]. Table VIII summarizes video distillation methods.

The main challenge in deep learning models, particularly in applications such as video quality estimation [445] and attention prediction [449], is the high computational cost and complexity of large models, which restrict their execution on edge devices. KD effectively addresses this by enabling a smaller, more efficient model (student) to learn from a larger, complex model (teacher) with minimal performance reduction while significantly decreasing computational and memory requirements. MobileVOS [442] specifically focuses on semi-supervised video object segmentation on resource-constrained devices, for example mobile phones, and proposes a pixelwise representation distillation loss to efficiently transfer structural information from the teacher to the student while ensuring feature consistency across frames. In [437], the authors propose a real-time approach that distills similarity information from an offline teacher model, which processes entire video sequences, to an online model, which processes individual frames. [450] introduces an online approach that applies feature distillation on live videos for low-cost semantic segmentation.

Furthermore, KD is applicable in multi-modal learning, enabling cross-modal knowledge transfer, such as video-to-image [433], video-to-text [446], text-to-video [438], audio-to-video [447], and video-to-audio [448]. Notably, [451] leverages KD in a spatio-temporal graph model for the video captioning task, making the model more robust against spurious correlations.

V2I-DETR enhances medical video lesion detection by distilling temporal knowledge from a video-based teacher model to an image-based student model, achieving real-time inference speed with high accuracy [433].

VI. APPLICATIONS

Knowledge distillation can be applied across various fields and has important applications, including distillation in LLMs, distillation in foundation models and in vision transformers, distillation in self-supervised learning, distillation in diffusion models, and distillation in visual recognition tasks. The following provides a detailed explanation of each application.

A. Distillation in Large Language Models

LLMs have recently revolutionized the field of NLP, with significant applications across various domains in both academia and industry. However, these models pose critical challenges due to their large number of parameters, which limits their deployment in real-time scenarios. Additionally, these giant models, with a considerable number of attention blocks, have been shown to be overparameterized [3], [4]. Therefore, KD plays a crucial role in compressing LLMs into Small Language Models (SLMs). Existing KD methods for LLMs are mainly classified into two major categories: white-box distillation and black-box distillation. In white-box KD, logits and intermediate outputs of the LLMs are accessible, enabling the student to benefit from the internal information of the teacher. This information can be either logits [10] or intermediate features [22], such as outputs of each attention block or attention scores. On the other hand, in black-box distillation, only the predictions of the teacher are available for distillation, which is common in many closed-source LLMs like GPT-4 [487] and Gemini [488], limiting the distillation process to only the predictions. Black-box methods are categorized into In-Context Learning (ICL) [484], Chain of Thought (CoT) [466], and Instruction Following (IF) [478]. Figure 8 summarizes the black-box and white-box distillation methods for LLMs.

Most white-box KD methods focus on the BERT [2] model [452]–[459]. DistilBERT [453] was one of the works that initialized the student with a pre-trained teacher and minimized the soft probabilities between the student and the teacher during the pre-training phase. In a concurrent work, MiniLM [457] proposed distilling the self-attention module of the last transformer layer of the teacher. MobileBERT [456] first trains a special teacher and then uses feature and attention transfer to distill knowledge to the student. TinyBERT [452] introduced a KD method for both pre-training and task-specific training of a smaller version of BERT by distilling the logits, attention matrices, hidden states, and the output of the teacher's embedding layer. PKD [454] introduces two strategies for incremental distillation, while [458] and [459] propose architecture-agnostic and task-agnostic distillation methods, respectively.

Recently, several new white-box KD methods have been proposed [460]–[465], primarily designed for decoder-based LLMs. MetaDistill [460] utilizes a feedback mechanism within a meta-learning framework, GKD [462] trains the student

Distillation in Large Language Models

- **White-box**
  - **BERT-based**: TinyBERT [452], DistillBERT [453], PKD [454], Turc et al. [455], MobileBERT [456], Minilm [457], Xtremdistil [458], Xtremedistiltransformer [459], MetaDistill [460]
  - **Decoder-based**: TED [461], GKD [462], Agarwal et al. [463], Pc-LoRA [464], Minillm [465]
- **Black-box**
  - **Chain of Thought**: Li et al. [466], Magister et al. [467], Fine-tune-CoT [468], MCC-KD [469], Distilling step-by-step [470], CSoTD [471], SCOTT [472], DOCTOR [473], PaD [474], TGD [475], RevThink [476], MiniPLM [477]
  - **Instruction Following**: Self-instruct [478], UniversalNER [479], LaMini-LM [480], Lion [481], Li et al. [482]
  - **In-context Learning**: Brown et al. [483], ILD [484], LLM-R [485], AICD [486]

Fig. 8. Classification of distillation methods for large language models.

on its self-generated output sequences, and PC-LoRA [464] concurrently conducts progressive model compression and fine-tuning. One of the most notable recent works is MiniLLM [465], which suggests using the reverse of the Kullback-Leibler Divergence to prevent student models from overestimating the low-probability distribution of the teacher.

Black-box KD has recently garnered increased attention due to the rise of closed-source LLMs. The first category of black-box KD is ICL, initially introduced in GPT-3 [483]. In ICL, the teacher receives a task description, a few task examples, and a query, where the teacher predicts a response that the student aims to mimic. [484] combines in-context learning objectives with language modeling objectives to distill both the ability to interpret in-context examples and task knowledge into smaller models. AICD [486] explores the simultaneous distillation of in-context learning and reasoning, leveraging the autoregressive nature of LLMs to optimize the likelihood of all rationales in context. [485] introduces a novel framework to iteratively train dense retrievers capable of identifying high-quality in-context examples for LLMs.

The second category of black-box KD is IF, which seeks to enhance the zero-shot capabilities of LLMs through instruction prompts. Specifically, in IF, the teacher generates task-specific instructions on which the student network is fine-tuned [478]–[482]. Self-Instruct [478] generates instructions, input, and output samples, filters out invalid samples, and then fine-tunes the student model. UniversalNer [479] explores mission-focused instruction tuning to train student models capable of excelling in a broad application class. Lion [481] utilizes an adversarial framework to generate challenging instructions for students, while LaMini-LM [480] compiles a large dataset containing 2.58 million instructions based on both existing and newly generated instructions.

The last and most popular black-box KD approach is COT, where the teacher model generates rationales alongside predictions, providing the student network with intermediate inference steps. In COT, the student is expected to generate predictions and reasonings similar to the teacher, similar to feature learning where the student mimics the teacher's intermediate steps for better predictions. This concept was first introduced by [466] and has been expanded upon in subsequent works [467]–[477]. [467] explores the trade-off between model and dataset size, demonstrating that the reasoning ability of LLMs can be transferred to SLMs. MCC-KD [469] generates multiple reasonings for each sample and enforces consistency and diversity among them. [470] employs a multi-task learning framework to train the student with the teacher's reasonings. In contrast to other approaches that overlook reasonings with incorrect labels, [475] utilizes both positive and negative data. Most recently, [476] demonstrated that generating backward questions and reasoning, and training the student on both forward and backward reasoning, significantly enhances the generalization ability of the student.

In summary, white-box KD methods possess greater efficacy in transferring knowledge by granting the student access to the internal information of the teacher. However, they typically come with high computational costs when distilling knowledge during student training. Moreover, most powerful LLMs are currently not open-source, rendering the use of white-box KD infeasible for these models. On the other hand, black-box methods enable KD from any model. Nevertheless, as black-box methods lack access to the internal workings of the teacher and solely rely on training the student with data generated by the teacher, they may not encompass all possibilities and could exhibit lower generalizability. One potential strategy for leveraging white-box distillation could involve initially distilling knowledge from a closed-source model to an open-source one using black-box KD, and subsequently applying a white-box method for further distillation.

B. Foundation Model Distillation

Foundation Models (FMs) are large-scale, pre-trained architectures designed to generalize across multiple downstream

Fig. 9. Categorization of foundation models based on their type, and the tasks for which their knowledge is distilled.

tasks. KD serves as a critical technique in optimizing these models by transferring their knowledge into smaller, task-specific, or efficient student models. KD addresses challenges such as computational costs and deployment on resource-constrained devices, making it indispensable for scaling FMs' applications.

Prominent FMs, including VLMs such as CLIP [263], conversational agents such as ChatGPT [483], [489], [490], and generative models such as DALL-E [491], illustrate the diverse domains from which knowledge can be distilled using KD. KD is utilized to distill the complex, multi-modal representations from these models into simpler ones optimized for specific tasks. In addition, vision-specific models such as SAM [492] and DINO [493] serve as valuable sources of knowledge that can be distilled for various applications, including segmentation and unsupervised object discovery. Transformer-based models, including ViT [494], DETR [495], Swin Transformer [496], and DeIT [497] also provide foundational knowledge that can be transferred through KD to enhance performance in vision-related tasks.

To provide a comprehensive overview of these FMs, this section explores KD for each model individually. Figure 9 presents a structured overview of the topics covered in this section, outlining the different types of FMs and their respective applications. To complement this, Figure 10 illustrates the architectures of these prominent models, highlighting key points where data is commonly extracted and distilled for downstream tasks.

1) Vision-language Models and Knowledge Distillation:
VLMs bridge visual and textual data through joint represen-

tation learning, offering capabilities to align these data types effectively. KD plays a central role in refining these models [485], [498], [499], enabling their general-purpose knowledge to be adapted for more compact or task-specific models, enhancing scalability and reducing computational overhead.

CLIP (Contrastive Language-Image Pre-training) [263], one of the most significant VLMs, exemplifies this integration. Trained on paired image and text data, CLIP achieves remarkable generalization by aligning global visual and textual representations. Through KD, these capabilities are transferred into student models designed for downstream tasks, including several key areas:

a) Open-vocabulary Tasks: In conventional vision tasks such as semantic segmentation and object detection, a dataset-focused methodology is traditionally employed, where the most successful techniques depend on a training dataset that has been manually annotated for a predefined and narrow range of categories. However, the rise of advanced VLMs, namely CLIP, is driving a shift towards an open-world paradigm. These models are trained through a simple yet scalable process that aligns image-text pairs, leveraging broad, loosely descriptive captions that can be collected in vast amounts with little human intervention. In object detection, KD techniques [500]–[506], [506]–[517] distill CLIP's global visual-textual knowledge into models capable of detecting objects outside fixed vocabularies, while in semantic segmentation [501], [518]–[530], and part segmentation [531] KD aids in transferring CLIP's textual alignment to dense pixel-level tasks, enabling fine-grained scene understanding. Furthermore, [532] has explored open-vocabulary customization from CLIP

Fig. 10. Architectural overview of prominent foundation models, indicating key points for data extraction and distillation to downstream tasks.

through Data-Free Knowledge Distillation (DFKD), introducing a framework that enables CLIP model adaptation without requiring the original training data.

b) Weakly Supervised Semantic Segmentation: The utilization of CLIP has recently been extensively explored for weakly supervised semantic segmentation [533]–[540] where it enables the generation of segmentation masks with minimal labeled data, by leveraging its ability to align textual and visual representations within a shared embedding space. Through this approach, text prompts representing class labels guide the identification of relevant regions in an image by querying CLIP's pre-trained image-text alignment capabilities. The visual features extracted by CLIP's image encoder are matched to the semantic meaning encapsulated in text embeddings, producing coarse class activation maps that localize objects and regions associated with the given labels.

c) Prompt Learning: Prompt learning is a method that enables the use of a large pre-trained model, such as CLIP, for specific tasks without having to retrain the entire model. This approach suggests adjusting the model's representations for particular tasks by using learnable soft prompts, either text-based or visual, rather than relying on pre-designed hard prompts. Several methods [541]–[545] have been proposed to implement this approach.

d) Generation and Editing: Applications such as generation and editing of 2D [546]–[548] and 3D images [549]–[556], where text-guided modifications are required, also benefit significantly from CLIP-guided KD.

e) Other Applications: The range of tasks addressed by KD in VLMs, particularly CLIP, is broad, including person re-identification [557], zero-shot human-object interaction (HOI) detection [558], open-vocabulary out-of-distribution classification [559], domain generalization [560], video-language retrieval [561], affordance detection [562], video highlight detection [563], and monocular depth estimation [564]. These applications demonstrate how KD transforms large-scale VLMs into accessible, efficient tools for specialized needs, with some works [565]–[568] tackling multiple tasks simultaneously.

2) Segment Anything (SAM): The Segment Anything Model (SAM) [492] is a versatile and general-purpose model designed for efficient and accurate object segmentation across diverse image domains. Developed with a focus on flexibility, SAM employs a powerful vision-transformer architecture that can segment any object in an image based on prompts, such as points, boxes, or masks. Its pre-trained nature enables SAM to generalize well across tasks without task-specific fine-tuning, making it a valuable resource for many applications. Some works [323], [569], [570] leverage SAM's pre-trained knowledge by distilling its segmentation capabilities into lighter, task-specific models, enabling the transfer of its robust segmentation performance to applications with constrained data annotations, computational resources or domain-specific requirements.

a) SAM for Weakly Supervised Semantic Segmentation: SAM is utilized for weakly supervised semantic segmentation [571]–[575] by leveraging its pre-trained capabilities to

generate high-quality pseudo-labels from minimal or weak supervision signals, such as image-level tags, points, or bounding boxes. These pseudo-labels serve as a foundation for training more efficient models, bridging the gap between limited supervision and fully labeled data. The robustness of SAM's outputs allows semantic segmentation models to learn complex patterns and object boundaries without extensive manual annotation. This approach significantly reduces labeling costs while maintaining competitive performance.

3) Multi-model Applications: In some instances, research has explored synergistic use of multiple FMs, combining their strengths through KD to enhance performance in segmentation, detection, and multi-modal applications. This collaborative use of KD fosters the development of unified models that inherit the complementary strengths of various teacher architectures. General studies have demonstrated the effectiveness of integrating diverse FMs across tasks [576]–[583], highlighting improvements in both efficiency and accuracy. One common approach involves combining models with complementary strengths. For instance, (SAM + CLIP) [584]–[586] leverages SAM's robust image segmentation alongside CLIP's semantic understanding, enhancing performance in various vision tasks. Similarly, (DINO + CLIP) [587], [588] has been explored, taking advantage of DINO's self-supervised learning capabilities and CLIP's strong vision-language alignment. Extending this idea, (DETR + CLIP) has been proposed in [589] for open-vocabulary object detection, where DETR serves as the detection model while CLIP provides text-based prompts. Additionally, research [590], [591] has investigated using the complementary strengths of (SAM + GDINO) for object detection and segmentation.

4) Vision Transformers: Vision Transformers (ViTs) leverage the self-attention mechanism to capture long-range dependencies in images, enabling them to model complex and global visual features. This capability makes them highly effective for tasks such as image classification, segmentation, and detection. However, the high computational cost of training and deploying ViTs highlights the need for distilling the knowledge of pre-trained transformer models for more specialized tasks. The following section discusses prominent ViTs and their distillation methods.

a) Vision Transformer: ViT [494] is a deep learning model that extends the Transformer architecture [592], originally developed for natural language processing, to image analysis. Several works [593]–[599] have explored the application of KD to ViT, investigating a range of strategies aimed at achieving diverse objectives, as summarized in Table IX.

b) Distillation with No Labels (DINO): DINO [493] is a self-supervised vision transformer that learns rich visual representations from unlabeled data. Due to its strong ability to capture meaningful features, many works leverage Dino's pre-trained features for object localization for various downstream tasks, such as unsupervised object discovery [478], [600]–[604], open-world counting [605], and zero-shot semantic segmentation [606].

c) DEtection TRansformer (DETR): DETR [495] is an end-to-end object detection model built on ViT, capable of directly predicting bounding boxes and class labels within a

TABLE IX
SUMMARY OF METHODS APPLYING KNOWLEDGE DISTILLATION TO THE VISION TRANSFORMER MODEL.

Method	Contribution
SMKD [593] (2023)	Explores the tendency of ViTs to overfit and degrade in performance under few-shot learning, due to the lack of CNN-like inductive biases, and proposes a KD method to address this issue.
TinyMIM [594] (2023)	Explores KD techniques for transferring the benefits of large MIM (Masked Image Modeling) pre-trained models to smaller ones.
Incrementer [595] (2023)	Introduces a class-incremental semantic segmentation framework that employs ViT instead of traditional CNNs, emphasizing the need for global context modeling and incorporating convolutional layers to support new class predictions in incremental learning settings.
CST [596] (2023)	Uses self-distillation by leveraging attention maps from a frozen ViT to generate pseudo-labels for few-shot, weakly supervised tasks.
DMAE [597] (2023)	Suggests distilling intermediate features, rather than logits, from a large ViT model to a smaller one.
CSKD [598] (2023)	Presents a method for enhancing ViT performance through KD from CNN models.
ViTKD [599] (2024)	Introduces a method for distilling knowledge between ViT feature maps instead of logits, using DeiT [497] as the base model.

unified framework, thereby eliminating the need for anchors or complex post-processing. Its ability to effectively capture global context and model interactions between objects has positioned it as a leading backbone in object detection tasks. However, DETR's large model size and substantial computational demands pose significant challenges for deployment in real-world scenarios with limited computational budgets. To overcome these limitations, recent works [607], [608] have explored compressing DETR through KD, aiming to retain its strong performance while reducing computational overhead for real-time and resource-limited applications.

d) Swin Transformer: Swin Transformer [496] is a powerful architecture that effectively combines the strengths of CNNs and ViTs through its hierarchical design and efficient window-based attention mechanism. By employing local window attention within each window and shifting the windows between layers, Swin captures both local and global context, enabling it to excel in modeling fine-grained details as well as long-range dependencies. This rich feature representation has made Swin a preferred choice in various computer vision tasks. Recent works [609]–[621] have leveraged Swin Transformer as an encoder, utilizing its features across different vision tasks, as summarized in Table X.

C. Distillation in Self-supervised learning

Self-Supervised Learning (SSL) is an important field due to its independence from labeled data, and self-distillation is the basis of the most prominent works in SSL, especially in the realm of computer vision. The most prominent work in SSL was based on self-distillation [622]. SimCLR [622] proposed a shared network that receives two different augmentations of an image and minimizes their extracted embeddings. Following SimCLR, different extensions have been proposed [623]–[628]. MOCO [623], [624] and BYOL [625] are among successful methods where they propose to use the same network

TABLE X
SUMMARY OF METHODS APPLYING KNOWLEDGE DISTILLATION TO THE
SWIN TRANSFORMER MODEL.

Method	Task
MaskFormer [611] (2021)	Semantic Segmentation
DFR [610] (2021)	Weakly Supervised Semantic Segmentation (Scribbles)
SwinNet [612] (2021)	Salient Object Detection
SwinTrack [613] (2022)	Siamese Tracking
SwinF [615] (2022)	Target Detection
Swin-UNETR [616] (2022)	Medical Image Analysis
Ssformer [614] (2022)	Semantic Segmentation
HGI-SAM [617] (2023)	Weakly Supervised Intracranial Hemorrhage Segmentation
HiFormer [618] (2023)	Medical Image Segmentation
Swin-Fusion [619] (2023)	Human Action Recognition
P.Swin [620] (2023)	Image Classification and Object Detection
SCUNet++ [621] (2024)	Pulmonary Embolism CT Image Segmentation
SWTformer [609] (2024)	Weakly Supervised Semantic Segmentation (Class Labels)

but with a different training process. The student is trained using backpropagation, and the teacher is an EMA of the student. SwAP [627] uses some prototypes and assigns a code to the augmentations of the same image. The student is then trained to predict the same class ID as the teacher. DINO [493] and DINOv2 [629] are the most recent important applications of KD in SSL, where global and local patches of an image pass through similar teacher and student networks. The teacher is updated using EMA, and the representations of the teacher and the student are forced to be similar. This local-to-global approach makes it ideal for segmentation tasks.

Recently, more complex forms of SSL based on KD have been proposed [444], [626], [628], [630]–[632]. BINGO [628] uses a pre-trained SSL method to group the images of a dataset into a bag of samples. Then, in each epoch, samples are taken from the bag and the inter and intra-sample losses are minimized. [444] trains two SSL teachers for images and videos using masked modeling and follows a scenario similar to SimCLR, minimizing the spatial features with the image teacher and the spatio-temporal features with the video teacher. [630] uses a parallel SSL approach and defines consistency losses between them, and [631] integrates a known KD approach into an SSL framework using an SSL pre-trained teacher.

Apart from the mentioned papers, some studies attempt to distill the knowledge of a pre-trained SSL network into a smaller network [633]–[637]. The gap between two similar networks, one trained fully-supervised and one trained self-supervised, is shown to decrease with a bigger model size [633]. [635] shows the impact of an MLP layer after representations of the student, while [633] proposes a similarity-based distillation method for distilling the knowledge of a pre-trained network, which can outperform the corresponding fully-supervised trained network. In a similar approach, SEED [636] uses a queue of samples and minimizes the distance of the student's representation with anchors in the queue and the corresponding distances between the anchors and the teacher's representation. [637] proposes a novel approach to complement self-supervised pretraining via an auxiliary pretraining phase for small unlabeled datasets.

D. Diffusion Distillation

Diffusion models [73], [638] generate high-quality images, text, and 3D data through an iterative denoising process that can require hundreds or thousands of steps. This is computationally expensive, limiting deployment on low-resource hardware. KD mitigates this by transferring soft outputs and internal representations from a large teacher model to a smaller student model. The distilled model then achieves the same quality with far fewer sampling steps, requiring significantly less inference time and computation [639].

Several strategies have been developed to accelerate diffusion models. Feature-level methods have the student mimic the teacher's internal representations [72], [640]–[642]. Sampling process distillation reduces the iterative denoising by either merging multiple steps into one or matching a one-step student's output to that of the full teacher [643], [644]. Adversarial and data-free approaches align outputs using GAN losses or bootstrapping without requiring large synthetic datasets [160], [645], [646]. Consistency models reformulate the denoising into a direct, often single-step transformation, cutting down inference time [647]–[652]. In the following sections, each category is explained in detail.

Feature-level: Feature-level distillation accelerates diffusion model efficiency by bringing the student representations closer to the teacher and reducing inference steps. Recent methods reinforce feature alignment through denoising [72], external guidance [640], and cross-sample consistency [641], which facilitate convergence and improved sampling efficiency. Furthermore, combining feature distillation with model compression reduces model size and inference time [642]. These techniques accelerate generation without compromising output quality, thereby making diffusion models more realistic.

Sampling Process: Distillation of the sampling procedure is performed to ease the naturally iterative denoising process characteristic of diffusion models. Progressive distillation [643] achieves this by consolidating two steps into a single step; through repeated application of this simplification, the duration of the sampling procedure can be drastically shortened. Distribution matching distillation [644] attempts to condense the entire iterative procedure into a single step. This alignment is accomplished by matching the output distribution of a one-step student model directly with that of the combined multi-step teacher model.

Adversarial and Data-free: Adversarial methods [160], [645] leverage GAN-based losses to enforce the student's output distribution to closely match that of the teacher. Alternatively, data-free methods like EM distillation [646], employ bootstrapping between successive denoising steps, eliminating the need for large synthetic datasets during training.

Consistency Models: Consistency models [647]–[649] reformulate the denoising process as a direct, often single transformation from noisy data to its clean version. Their latent-space variants [650], [651], along with improvements focusing on stability and scalability [652], achieve significant speedups in inference times. However, training these models may require careful tuning to prevent quality collapse.

Furthermore, recent developments in the area of distillation techniques expanded their accessibility to multi-modal and multi-model settings, thus improving transferability across tasks and diverse architectures. ProlificDreamer [653] and DREAMFUSION [654] utilize the methods in the realm of text-to-3D translation, and general-purpose systems like Diff-Instruct [655] can facilitate generalization for the task.

In summary, various approaches offer distinct benefits: Sampling process distillation drastically accelerates generation but tends to require cumbersome training in order to maintain sample quality. Adversarial and data-free methods provide more flexibility, though sometimes at the cost of training stability. Consistency models enable low-latency generation, provided they are well-calibrated. Furthermore, combining cross-modal and universal approaches significantly expands the applicability of these distillation techniques across a wide range of generation tasks.

E. Visual Recognition Distillation

Knowledge distillation has been widely applied to computer vision tasks, enabling the training of compact student models replicating the behavior of larger, resource-intensive teacher models. Key applications of KD in visual recognition are summarized in Table XI, based on two main factors: the type of knowledge transferred (feature-based, logit-based, similarity-based, and combinations) and the distillation scheme employed (offline, online, and self-distillation). Several important visual recognition tasks are covered in this table, including depth estimation [656]–[671], face recognition [672]–[683], image segmentation [40], [54], [57], [59], [67], [70], [76], [77], [196], [251], [595], [684]–[695], medical image analysis [265], [696]–[716], object detection [70], [507], [686], [717]–[744], object tracking [745]–[748], pose estimation [93], [749]–[760], super resolution [262], [761]–[773], action recognition [214], [431], [774]–[780], and image retrieval [781]–[790].

VII. Performance Comparison

KD enables the compression of a deep network into a smaller model while preserving strong performance. Due to the vast number of methods proposed in this field, a comprehensive comparison across existing methods in different settings is necessary. However, the complicated nature of KD, encompassing various schemes, algorithms, sources, and teacher-student architectures, makes it challenging to conduct a comprehensive and fair comparison of all existing methods. For example, the exact pre-trained teacher model along with factors such as the number of epochs and batch size, can influence the results of the distillation. These factors are highly dependent on the hardware resources employed in each method, which is the determining factor in the final results.

In this work, existing methods are compared for two well-known tasks: classification and semantic segmentation segmentation, as almost all the proposed distillation methods evaluate their results on these two tasks. For classification, the CIFAR-100 dataset, and for semantic segmentation, the Cityscapes dataset are used, as these two datasets are widely employed in existing methods. CIFAR-100 is composed of

32×32 images taken from 100 classes, with 50,000/10,000 images for training/testing, and each class has the same number of training and testing images. Cityscapes is designed for understanding urban scenes and includes 2,975/500/1,525 images for training/validation/testing in 19 classes.

Furthermore, due to the importance of distillation in LLMs nowadays, compare existing methods for distillation in LLMs are compared. This includes comparing black-box and white-box methods. However, datasets and models used in LLMs vary significantly across different papers, specifically in white-box methods. Results for the datasets and models that are more commonly used in existing methods are reported, making the comparison easier and fair. The datasets used are: GLUE, Multilingual NER, DollyEval, IMDB, GSM8K, Cross-Fit, CommonsenseQA, SVAMP, Universal NER Benchmark, OpenBookQA, BBH, StrategyQA, and MATH.

Table XII summarizes the distillation methods for classification and segmentation tasks. For each method, the knowledge source, teacher and student architectures, and accuracy (for classification) and mIoU (for segmentation) after distillation are reported. For a fair comparison, numbers are directly reported from the corresponding paper. However, in some cases where the original paper does not provide results on the mentioned datasets, and direct comparison is not possible, results obtained from implementing their models are reported (denoted by *).

In a similar vein, Table XIII presents a comparison of distillation methods for LLMs. It shows the compression rate and the improvement over the teacher model for each method. As can be seen, distillation in LLMs is more crucial than in other fields, as it allows for compressing large models with a high compression rate while still maintaining comparable performance. It should be noted that in cases where the exact performance of the teacher model is not reported, performance of the student model before and after distillation is reported (denoted by *).

VIII. Conclusion and Discussion

Knowledge distillation has revolutionized the field of model compression and has played a significant role in various research topics in recent years. In this work, a comprehensive survey on knowledge distillation was proposed, reviewing its methods from various perspectives, including: sources, schemes, algorithms, modalities, applications, and performance comparisons. In contrast to existing surveys, this survey presented recent advancements in the field and focuses on the most important topics in KD. Adaptive and contrastive distillation were presented as two important algorithms that have gained increased attention in recent years. Additionally, KD was reviewed across different modalities, particularly for 3D inputs such as point clouds, which have established important connections with KD in recent research. Furthermore, the applications of distillation in self-supervised learning, diffusion models, foundation models, and LLMs was explored. Self-supervised learning methods primarily rely on self-distillation techniques, diffusion models utilize distillation to reduce the number of steps in the generation phase, and

TABLE XI
SUMMARY OF KNOWLEDGE DISTILLATION METHODS IN VISUAL RECOGNITION TASKS BASED ON THEIR SOURCES AND SCHEMES OF DISTILLATION.

Task	Knowledge	Distillation Scheme	Reference
Depth Estimation	feature	offline	[656]
		self-distillation	[657]
	logit	offline	[658], [659], [660], [661], [662]
		self-distillation	[663], [664], [665], [666], [667], [668]
	similarity	offline	[669]
	feature + logit	offline	[670]
		online	[671]
Face Recognition	feature	offline	[672], [667], [673], [674], [675], [676], [677], [678]
	logit	offline	[679], [680], [681]
	similarity	offline	[682]
		self-distillation	[683]
Image Segmentation	feature	offline	[684], [70], [685], [595], [77], [76], [67]
		self-distillation	[686]
	logit	offline	[196], [659], [687]
		online	[688], [689]
	similarity	offline	[59], [57]
		online	[690]
	feature + logit	offline	[251], [691], [692], [693]
		online	[694]
	feature + similarity	offline	[54], [40]
		online	[695]
Medical Image Analysis	feature	offline	[696], [697], [698], [699], [700], [701], [702], [703]
		online	[704]
		self-distillation	[705], [706]
	logit	offline	[707], [708]
		online	[709]
		self-distillation	[710], [711]
	similarity	online	[265]
		self-distillation	[712], [713]
	feature + logit	offline	[714]
	feature + similarity	offline	[715], [716]
Object Detection	feature	offline	[717], [718], [719], [720], [721], [722], [70], [723], [724], [507], [725]
		self-distillation	[726], [727], [686], [728]
	logit	offline	[729], [730], [731]
		self-distillation	[732]
		online	[733]
	similarity	offline	[734]
	feature + logit	offline	[735], [736]
		online	[737], [738], [739], [740], [741], [742]
	feature + similarity	offline	[737], [738], [739], [740], [741], [742]
		self-distillation	[743]
	logit + similarity	offline	[744]
Object Tracking	feature	offline	[745], [746]
	logit	offline	[747]
	feature + similarity	offline	[748]
Pose Estimation	feature	offline	[749]
		self-distillation	[750]
	logit	offline	[751], [752], [753], [754], [755], [756]
		online	[93]
		self-distillation	[757]
	feature + logit	offline	[758], [759]
		self-distillation	[760]
Super Resolution	feature	offline	[761], [262]
		self-distillation	[762], [763]
	logit	offline	[764]
		self-distillation	[765]
	relation	offline	[766], [767]
	feature + similarity	offline	[768], [769]
		self-distillation	[770]
	feature + logit	offline	[771], [772], [773]
Action Recognition	feature	offline	[774], [431], [214], [775], [776]
		online	[777]
	logit	offline	[778]
	feature + similarity	offline	[779], [780]
Image Retrieval	feature	offline	[781], [782], [783]
	similarity	offline	[784], [785], [786], [787]
	feature + logit	offline	[788]
	feature + similarity	offline	[789], [790]

TABLE XII
PERFORMANCE COMPARISON OF DIFFERENT KNOWLEDGE DISTILLATION METHODS ON CIFAR-100 AND CITYSCAPES DATASETS (* DENOTES THAT RESULTS ARE NOT REPORTED FROM THE ORIGINAL PAPER).

Method	Knowledge	Teacher (baseline)	Student (baseline)	Accuracy	Code
Classification (CIFAR-100)					
FitNet [22] (2014)	Feature	WRN-40-2 (75.61)	WRN-40-1 (71.98)	72.24 (+0.26)*	Link
KD [10] (2015)	Logit	WRN-40-2 (75.61)	WRN-40-1 (71.98)	73.54 (+1.56)*	Link
AT [23] (2016)	Feature	WRN-40-2 (75.61)	WRN-40-1 (71.98)	72.77 (+1.79)*	Link
CCKD [86] (2019)	Similarity	ResNet-110 (–)	ResNet-20 (68.40)	72.40 (+4.00)	–
SPKD [25] (2019)	Similarity	WRN-40-2 (75.61)	WRN-40-1 (71.98)	72.43 (+1.45)*	–
Chen et al. [48] (2020)	Similarity + Logit	ResNet-101 (71.77)	ResNet-18 (65.64)	69.14 (+3.49)	Link
TAKD [18] (2020)	Logit	ResNet-110 (–)	ResNet-8 (–)	61.82	Link
SphericalKD [29] (2020)	Logit	ResNet-32×4 (79.42)	ResNet-8×4 (72.50)	76.40 (+3.90)	Link
RKD [68] (2021)	Feature	WRN-40-2 (75.61)	WRN-40-1 (71.98)	75.09 (+3.11)	Link
SemCKD [69] (2021)	Feature + Logit	WRN-40-2 (75.61)	MobileNetV2 (65.43)	69.67 (+4.24)	Link
ICKD [24] (2021)	Similarity + Logit	WRN-40-2 (75.61)	WRN-40-1 (71.98)	74.63 (+2.65)	Link
DKD [35] (2022)	Logit	WRN-40-2 (75.61)	WRN-40-1 (71.98)	74.81 (+2.83)	Link
DistKD [55] (2022)	Similarity + Logit	WRN-40-2 (75.61)	WRN-40-1 (71.98)	74.73 (+2.75)	Link
Norm [79] (2023)	Feature	WRN-40-2 (75.61)	WRN-40-1 (71.98)	74.82 (+2.84)	Link
SFKD [34] (2024)	Logit	ResNet-32×4 (79.55)	ResNet-8×4 (72.50)	76.84 (+4.34)	–
NormKD [30] (2024)	Logit	ResNet-32×4 (79.42)	ResNet-8×4 (72.50)	76.57 (+4.07)	Link
CRLD [63] (2024)	Similarity + Logit	WRN-40-2 (75.61)	WRN-40-1 (71.98)	75.58 (+3.60)	Link
NTCE-KD [36] (2024)	Logit	WRN-40-2 (75.61)	WRN-40-1 (71.98)	76.44 (+4.46)	–
TTM [32] (2024)	Logit	WRN-40-2 (75.61)	WRN-40-1 (71.98)	74.58 (+2.60)	Link
Miles et al. [31] (2024)	Logit	WRN-40-2 (76.46)	WRN-16-2 (73.64)	76.14 (+2.50)	Link
SDD [39] (2024)	Logit	ResNet-32×4 (79.42)	MobileNetV2 (64.6)	68.84 (+4.24)	Link
Sun et al. [20] (2024)	Logit	WRN-40-2 (75.61)	ResNet-8×4 (72.50)	77.11 (+3.14)	Link

Method	Knowledge	Teacher (baseline)	Student (baseline)	mIoU	Code
Segmentation (CityScapes)					
KD [10] (2015)	Logit	ResNet-101 (77.66)	ResNet-18 (64.09)	65.21 (+1.12)*	Link
AT [23] (2016)	Feature	ResNet-101 (77.66)	ResNet-18 (64.09)	65.29 (+1.20)*	Link
Xie et al. [84] (2018)	Similarity + Logit	ResNet-101 (70.90)	MobileNetV2 (67.30)	71.90 (+4.60)	–
SKD [58] (2019)	Feature + Logit	ResNet-101 (78.40)	MobileNetV2 (67.60)	71.40 (+3.80)	Link
He et al. [82] (2019)	Feature + Similarity	ResNet-101 (71.40)	MobileNetV2 (70.20)	72.70 (+2.50)	–
IFVD [26] (2020)	Similarity + Logit	ResNet-101 (78.56)	ResNet-18 (69.10)	74.54 (+5.44)	Link
CWD [67] (2021)	Feature	ResNet-101 (78.50)	ResNet-18 (63.63)	71.03 (+7.40)	Link
DSD [54] (2021)	Similarity	ResNet-101 (78.23)	ResNet-18 (69.42)	72.24 (+2.82)	–
CIRKD [59] (2022)	Similarity + Logit	ResNet-101 (78.07)	ResNet-18 (72.55)	74.73 (+2.18)	Link
DIST [55] (2022)	Similarity + Logit	ResNet-101 (78.07)	ResNet-18 (72.55)	76.31 (+3.76)	Link
IDD [56] (2022)	Similarity	ResNet-101 (78.40)	MobileNetV2 (67.60)	76.33 (+8.73)	–
MGD [70] (2022)	Feature	ResNet-101 (78.34)	ResNet-18 (73.20)	76.31 (+3.11)	Link
MLP [71] (2023)	Feature	ResNet-101 (78.34)	ResNet-18 (73.20)	76.55 (+3.35)	–
DiffKD [72] (2023)	Feature	ResNet-101 (78.07)	ResNet-18 (74.21)	77.78 (+3.57)	Link
FAKD [77] (2024)	Feature	ResNet-101 (79.74)	ResNet-18 (68.99)	74.75 (+5.76)	–
BPKD [40] (2024)	Logit	ResNet-101 (79.74)	ResNet-18 (74.23)	77.57 (+3.34)	Link
LAD [76] (2024)	Feature + Logit	ResNet-101 (79.76)	ResNet-18 (72.65)	76.86 (+4.21)	–
FreeKD [78] (2024)	Feature	ResNet-101 (70.43)	ResNet-18 (73.20)	76.45 (+3.25)	Link
AttnFD [74] (2024)	Feature	ResNet-101 (77.66)	ResNet-18 (64.09)	73.04 (+8.96)	Link
AICSD [57] (2025)	Similarity + Logit	ResNet-101 (77.66)	ResNet-18 (64.09)	70.96 (+6.88)	Link

TABLE XIII
PERFORMANCE COMPARISON OF DIFFERENT BLACK-BOX AND WHITE-BOX DISTILLATION METHODS FOR LLMs (* DENOTES THAT THE COMPARISON IS BETWEEN THE STUDENT MODEL BEFORE AND AFTER DISTILLATION).

Method	Distillation Type	Teacher Model	Compression Rate	Benchmark	Comparison with Teacher		Code
White-box Distillation							
DistillBERT [453] (2019)	Logit	$BERT_{base}$	1.66	GLUE	77.00 / 79.50	97.0%	–
TinyBERT [452] (2019)	Feature	$BERT_{base}$	7.50	GLUE	77.00 / 79.50	97.0%	Link
PKD [454] (2019)	Logit + Feature	$BERT_{base}$	2.00	GLUE	77.70 / 84.90	92.0%	Link
PD [455] (2019)	Logit	$BERT_{base}$	2.00	GLUE	82.10 / 81.70	100.5%	Link
Xtremedistil [458] (2020)	Feature	$mBERT_{base}$	35.00	Multilingual NER	88.60 / 92.70	95.0%	Link
MobileBERT [456] (2020)	Feature	$BERT_{base}$	4.30	GLUE	77.70 / 78.30	99.0%	Link
MINILM [457] (2020)	Feature	$BERT_{base}$	1.65	GLUE	80.40 / 81.50	98.0%	–
xtremedistiltransformers [459] (2021)	Feature	$BERT_{base}$	7.70	GLUE	81.70 / 83.70	97.6%	Link
MetaDistil [460] (2021)	Feature	$BERT_{base}$	2.00	GLUE	80.40 / 80.70	99.0%	Link
TED [461] (2023)	Feature	$DeBERTaV3_{base}$	2.60	GLUE	87.50 / 88.90	98.0%	Link
MiniLLM [465] (2024)	Feature	GPT-2	2.00	DollyEval	57.40 / 58.40	94.0%	Link
PC-LoRA [464] (2024)	Feature	$BERT_{base}$	3.73	IMDB	92.78 / 94.00	98.0%	–
GKD [463] (2024)	Logit	$T5_{XL}$	3.75	GSM8K	29.10 / 29.00	100.3%	–
Black-box Distillation							
ILD [484] (2022)	ICL	$GPT2_{large}$	6.00	CrossFit	61.20 / 66.20	92.0%	–
LLM-R [485] (2023)	ICL	$LLaMA_{13B}$	2.00	CommonsenseQA	48.80 / 49.60	98.0%	–
AICD [486] (2024)	ICL	GPT-3.5-Turbo	–	SVAMP	51.70 / 20.70	250.0%*	–
UniversalNER [479] (2023)	IF	GPT-3.5-Turbo	–	UNIVERSAL NER	41.70 / 34.90	119.0%	Link
LaMini-LM [480] (2023)	IF	$Alpaca_{7B}$	9.00	OpenBookQA	34.00 / 43.20	79.0%	Link
Lion [481] (2023)	IF	GPT_{175B}	25.00	BBH	32.00 / 48.90	65.0%	Link
Fine-tune-CoT [468] (2022)	CoT	GPT_{175B}	26.00	SVAMP	30.33 / 64.67	47.0%	Link
Li et al. [466] (2022)	CoT	GPT_{175B}	58.00	CommonsenseQA	82.47 / 73.00	113.0%	–
Magister et al. [467] (2022)	CoT	GPT_{175B}	16.00	StrategyQA	63.77 / 65.40	97.0%	–
MCC-KD [469] (2023)	CoT	GPT-3.5-Turbo	–	SVAMP	68.66 / 75.14	91.0%	Link
CSoTD [471] (2023)	CoT	GPT_{175B}	135.00	CommonsenseQA	67.00 / 82.10	82.0%	–
SCOTT [472] (2023)	CoT	$GPT neox_{20B}$	7.00	CommonsenseQA	74.70 / 60.40	124.0%	Link
Distilling step-by-step [470] (2023)	CoT	$Palm_{540B}$	700.00	SVAMP	58.40 / 72.30	81.0%	Link
PaD [474] (2023)	CoT	GPT-3.5-Turbo	–	SVAMP	36.70 / 57.80	64.0%	–
TDG [475] (2024)	CoT	GPT-3.5-Turbo	–	MATH	6.81 / 3.88	175.0%*	Link
RevThink [476] (2024)	CoT	Gemini-1.5-Pro-001	30.00	CommonsenseQA	75.76 / 76.72	99.0%	–

LLMs are distilled into smaller versions KD was reviewed across different modalities, particularly for 3D inputs such as point clouds, which have established important connections with KD in recent research. Furthermore, with recent advancements in foundation models like CLIP, distilling their rich knowledge into other models has become increasingly widespread.

In the following, the current challenges of existing distillation methods are discussed, and insights into future research directions in knowledge distillation are provided.

A. Challenges

While KD is an effective method for improving the performance of smaller networks, it comes with several challenges. Below, some of the most important challenges are discussed.

Knowledge Extraction: Finding an appropriate source of knowledge for distillation is one of the main challenges in the distillation process [66], [791], [792]. Although using logits is the most straightforward approach to distillation, its effectiveness varies across different scenarios. For example, in tasks where labels are not available, logit-based distillation is

not feasible. Furthermore, with the emergence of foundation models like CLIP, distilling rich features and embeddings has become crucial, which is not achievable through logit-based distillation alone, requiring feature-based and similarity-based distillation methods. In LLMs, this issue is even more pronounced, as some recent LLMs are not open-source [487], [488], restricting access to intermediate features or logits, and making black-box distillation particularly challenging. Existing methods often leverage a combination of different distillation sources to effectively transfer the teacher model's knowledge to the student model.

Choosing Proper Distillation Scheme: Selecting an appropriate teacher-student architecture is another challenge, as it depends on the specific context. In scenarios where a pre-trained large model is available, offline distillation tends to be more effective, whereas, in other cases, online distillation may prove to be a better alternative. In self-supervised learning, self-distillation schemes have shown to be highly effective. Additionally, using multiple teachers to train a smaller network is a promising approach, often referred to as a mixture of experts. However, this comes with its own challenges, including the need for more powerful memory and GPUs, as well

as the complexity of loading different teachers onto separate resources. In most existing methods, offline distillation is more commonly used. Specifically, when distilling from foundation models or LLMs, training the teacher alongside the student is not feasible due to the immense number of parameters in the teacher model.

Capacity Gap: Another important challenge is the capacity gap between the teacher and student models. It may be assumed that a larger teacher always leads to higher performance improvement in the student network, but when the teacher and student networks have a considerable size difference, it can cause capacity gap issues [29], [55], [88]. This is because the receptive fields of the teacher and student models differ significantly. This challenge is even more pronounced in foundation models, as they contain a tremendous number of parameters, making the distillation of their knowledge into a smaller model challenging. Furthermore, in black-box distillation of LLMs, where intermediate layers are not accessible, directly transferring knowledge from the teacher to the student is a serious challenge. Existing methods employ various algorithms to reduce this gap and make the teacher's knowledge more comprehensible for the student.

Architectural Difference: Differences in the architecture of the teacher and student models present another significant challenge, which can impact the performance of the distillation method. In scenarios where the teacher and student have different architectures, which is very common, transferring proper knowledge becomes a new challenge. This issue becomes apparent when distilling knowledge from a teacher to a student with a different architecture, often leading to performance degradation [29], [88]. In foundation models, the student's architecture can differ significantly from that of the teacher's. Although logit-based distillation remains effective in such scenarios, feature-based and similarity-based distillation methods face additional challenges. More specifically, selecting appropriate layers with similar semantic information for feature distillation plays an important role in the performance of the method, as the teacher and student may even have different inductive biases. In black-box distillation for LLMs, CoT distillation can help address this problem, as it does not force the student to mimic the exact intermediate layers of the teacher. Instead, it provides the student with intermediate reasoning steps derived from the teacher. However, generating proper reasoning and addressing the capacity gap between the teacher and student remain significant challenges.

B. Future Directions

Although KD is not a new concept and has been around since its initial proposal, the number of new methods being proposed is rapidly increasing, along with its expanding applications in various fields. This presents a significant opportunity for further research. In the following, potential future directions are discussed.

Feature-based Distillation: While logit-based and similarity-based distillation are effective, recent trends and results show that feature-based distillation can lead to even greater improvements in student performance. However, it can also be combined with other methods. Recent feature-based distillation approaches have demonstrated that instead of defining complex relationships, a simple transformation on the features of either the teacher or student is sufficient to achieve SOTA results in enhancing the student's performance [71], [76]. Furthermore, with the rapid growth of foundation models like CLIP, distilling embeddings from these large-scale models is gaining increasing attention.

Adaptive Distillation: Despite the existence of numerous distillation approaches, effectively distilling informative knowledge from the teacher to the student remains largely unexplored [791], [792]. Several adaptive distillation methods have been proposed to effectively transfer the dark knowledge of the teacher while ignoring unused knowledge. This line of research is particularly interesting, as it prevents the distillation of incorrect knowledge to the student and helps reduce the gap between the teacher and student networks. This process closely resembles human learning: a teacher transfers knowledge to students, but directly transferring all knowledge is not always effective or practical [18], [88]. Therefore, various forms of adaptive distillation can be employed, such as using a teacher assistant, decreasing the impact of distillation toward the end of training, or applying adaptive loss functions based on the importance of samples. All of these approaches align with real-world learning scenarios in human training processes.

Distillation from Foundation Models: With the emergence of a variety of foundation and multi-modal models, they have become a rich source of knowledge for distillation [793]. For example, CLIP, with its image and text encoder, can provide student models with rich embeddings, where text can be used to improve vision tasks such as weakly-supervised semantic segmentation, making it a hot topic in research [533]. Additionally, due to the general-purpose features of these large-scale models, they can be leveraged for open-vocabulary tasks [604], where the task is not limited to predefined classes. These interesting applications of distillation present great potential for future research.

Distillation in LLMs: While distillation is a highly effective approach in vision, text, and speech tasks, its application in LLMs is even more important. As stated earlier, LLMs have been shown to be overparameterized [3], [4] with billions and even trillions of parameters, far exceeding the size of models in other fields. These weight matrices are often sparse [5], making distillation crucial in reducing model size while maintaining performance. Additionally, most powerful models are not open-source [487], [488], increasing the importance of black-box distillation. CoT distillation is a particularly interesting technique for transferring reasoning capabilities from an LLM to an SLM. Unlike conventional distillation methods, CoT aims to enhance the generalization ability of the student model by teaching it how to reason. Results show that by distilling the knowledge of an LLM, the student model can achieve performance comparable to the teacher while significantly reducing model size. This is especially important because the teacher model may be too large to run on common GPUs, whereas the distilled student model can be efficiently deployed on common hardware, making advanced AI more accessible to to a wider audience.

Applications of Distillation: In addition to proposing novel distillation methods, distillation has been applied across various tasks, resulting in significant performance improvements. For example, distillation has been successfully applied in adversarial attacks [166], [794], mitigating backdoor attacks [795], [796], and anomaly detection [797]. Furthermore, the concept of distillation is utilized in dataset distillation [798], making it an interesting research direction. Another application of distillation is its integration with other compression methods, such as quantization or low-rank factorization.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, no. 2. Minneapolis, Minnesota, 2019.

[3] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[4] T. Fischer, C. Biemann *et al.*, "Large language models are overparameterized text encoders," *arXiv preprint arXiv:2410.14578*, 2024.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[6] A. G. Howard, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[7] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[8] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[9] C. Buciluundefined, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. Association for Computing Machinery, 2006, p. 535–541.

[10] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[11] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[12] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 3048–3068, 2021.

[13] A. Alkhulaifi, F. Alsahli, and I. Ahmad, "Knowledge distillation in deep learning and its applications," *PeerJ Computer Science*, vol. 7, p. e474, 2021.

[14] C. Yang, X. Yu, Z. An, and Y. Xu, "Categories of response-based, feature-based, and relation-based knowledge distillation," in *Advancements in Knowledge Distillation: Towards New Horizons of Intelligent Systems*. Springer, 2023, pp. 1–32.

[15] C. Hu, X. Li, D. Liu, H. Wu, X. Chen, J. Wang, and X. Liu, "Teacher-student architecture for knowledge distillation: A survey," *arXiv preprint arXiv:2308.04268*, 2023.

[16] A. Moslemi, A. Briskina, Z. Dang, and J. Li, "A survey on knowledge distillation: Recent advancements," *Machine Learning with Applications*, p. 100605, 2024.

[17] J. Ba and R. Caruana, "Do deep nets really need to be deep?" *Advances in neural information processing systems*, vol. 27, 2014.

[18] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5191–5198.

[19] Z. Hao, J. Guo, K. Han, H. Hu, C. Xu, and Y. Wang, "Revisit the power of vanilla knowledge distillation: from small scale to large scale," *Advances in Neural Information Processing Systems*, vol. 36, pp. 10 170–10 183, 2023.

[20] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao, "Logit standardization in knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 731–15 740.

[21] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3779–3787.

[22] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[23] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[24] L. Liu, Q. Huang, S. Lin, H. Xie, B. Wang, X. Chang, and X. Liang, "Exploring inter-channel correlation for diversity-preserved knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8271–8280.

[25] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1365–1374.

[26] Y. Wang, W. Zhou, T. Jiang, X. Bai, and Y. Xu, "Intra-class feature variation distillation for semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 346–362.

[27] L. Song, X. Gong, H. Zhou, J. Chen, Q. Zhang, D. Doermann, and J. Yuan, "Exploring the knowledge transferred by response-based teacher-student distillation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2704–2713.

[28] B. B. Sau and V. N. Balasubramanian, "Deep model compression: Distilling knowledge from noisy teachers," *arXiv preprint arXiv:1610.09650*, 2016.

[29] J. Guo, M. Chen, Y. Hu, C. Zhu, X. He, and D. Cai, "Reducing the teacher-student gap via spherical knowledge disitllation," *arXiv preprint arXiv:2010.07485*, 2020.

[30] Z. Chi, T. Zheng, H. Li, Z. Yang, B. Wu, B. Lin, and D. Cai, "Normkd: Normalized logits for knowledge distillation," *arXiv preprint arXiv:2308.00520*, 2023.

[31] R. Miles and K. Mikolajczyk, "Understanding the role of the projector in knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4233–4241.

[32] K. Zheng and E.-H. Yang, "Knowledge distillation based on transformed teacher matching," *arXiv preprint arXiv:2402.11148*, 2024.

[33] C. Li, G. Cheng, and J. Han, "Boosting knowledge distillation via intra-class logit distribution smoothing," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[34] M. Yuan, B. Lang, and F. Quan, "Student-friendly knowledge distillation," *Knowledge-Based Systems*, vol. 296, p. 111915, 2024.

[35] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.

[36] C. Li, X. Teng, Y. Ding, and L. Lan, "Ntce-kd: Non-target-class-enhanced knowledge distillation," *Sensors*, vol. 24, no. 11, p. 3617, 2024.

[37] Y. Ding, G. Yang, S. Yin, J. Zhang, X. Fang, and W. Yang, "Generous teacher: Good at distilling knowledge for student learning," *Image and Vision Computing*, vol. 150, p. 105199, 2024.

[38] J. Cui, Z. Tian, Z. Zhong, X. Qi, B. Yu, and H. Zhang, "Decoupled kullback-leibler divergence loss," *Advances in Neural Information Processing Systems*, vol. 37, pp. 74 461–74 486, 2024.

[39] S. Wei, C. Luo, and Y. Luo, "Scaled decoupled distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 975–15 983.

[40] L. Liu, Z. Wang, M. H. Phan, B. Zhang, J. Ge, and Y. Liu, "Bpkd: Boundary privileged knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1062–1072.

[41] S. W. Kim and H.-E. Kim, "Transferring knowledge to smaller network with class-distance loss," 2017.

[42] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019.

[43] Q. Ding, S. Wu, H. Sun, J. Guo, and S.-T. Xia, "Adaptive regularization of labels," *arXiv preprint arXiv:1908.05474*, 2019.

[44] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1285–1294.

[45] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[46] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. Duan, "Knowledge distillation via instance relationship graph," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7096–7104.

[47] H. Chen, Y. Wang, C. Xu, C. Xu, and D. Tao, "Learning student networks via feature embedding," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 25–35, 2020.

[48] Z. Chen, X. Zheng, H. Shen, Z. Zeng, Y. Zhou, and R. Zhao, "Improving knowledge distillation via category structure," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 2020, pp. 205–219.

[49] Q. Wang, W. Yu, L. Che, C. Liu, Z. Zhang, J. Gong, and P. Chen, "Similarity knowledge distillation with calibrated mask," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 5770–5774.

[50] C. Cheng, W. Gao, and X. Bian, "Knowledge distillation via inter- and intra-samples relation transferring," in *2024 20th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. IEEE, 2024, pp. 1–7.

[51] H. Hu, H. Zeng, Y. Xie, Y. Shi, J. Zhu, and J. Chen, "Global instance relation distillation for convolutional neural network compression," *Neural Computing and Applications*, vol. 36, no. 18, pp. 10 941–10 953, 2024.

[52] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133–4141.

[53] C. Wang, J. Zhong, Q. Dai, Y. Qi, Q. Yu, F. Shi, R. Li, X. Li, and B. Fang, "Channel correlation distillation for compact semantic segmentation," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 37, no. 03, p. 2350004, 2023.

[54] Y. Feng, X. Sun, W. Diao, J. Li, and X. Gao, "Double similarity distillation for semantic image segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 5363–5376, 2021.

[55] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 716–33 727, 2022.

[56] Z. Zhang, C. Zhou, and Z. Tu, "Distilling inter-class distance for semantic segmentation," *arXiv preprint arXiv:2205.03650*, 2022.

[57] A. M. Mansourian, R. Ahamdi, and S. Kasaei, "Aicsd: Adaptive inter-class similarity distillation for semantic segmentation," *Multimedia Tools and Applications*, pp. 1–20, 2025.

[58] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2604–2613.

[59] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 319–12 328.

[60] C. Wang, J. Zhong, Q. Dai, Y. Qi, R. Li, Q. Lei, B. Fang, and X. Li, "Prrd: Pixel-region relation distillation for efficient semantic segmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[61] Q. Wang, L. Liu, W. Yu, S. Chen, J. Gong, and P. Chen, "Bckd: Block-correlation knowledge distillation," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 3225–3229.

[62] X. Xin, H. Song, and J. Gou, "A new similarity-based relational knowledge distillation method," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 3535–3539.

[63] W. Zhang, D. Liu, W. Cai, and C. Ma, "Cross-view consistency regularisation for knowledge distillation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2011–2020.

[64] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *arXiv preprint arXiv:1707.01219*, 2017.

[65] J. Kim, S. Park, and N. Kwak, "Paraphrasing complex network: Network compression via factor transfer," *Advances in neural information processing systems*, vol. 31, 2018.

[66] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proceedings of the*

[67] IEEE/CVF international conference on computer vision, 2019, pp. 1921–1930.

[67] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5311–5320.

[68] Y. Chen, Y. Xian, A. Koepke, Y. Shan, and Z. Akata, "Distilling audio-visual knowledge by compositional contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7016–7025.

[69] D. Chen, J.-P. Mei, Y. Zhang, C. Wang, Z. Wang, Y. Feng, and C. Chen, "Cross-layer distillation with semantic calibration," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 7028–7036.

[70] Z. Yang, Z. Li, M. Shao, D. Shi, Z. Yuan, and C. Yuan, "Masked generative distillation," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 53–69.

[71] Z. Liu, Y. Wang, X. Chu, N. Dong, S. Qi, and H. Ling, "A simple and generic framework for feature distillation via channel-wise transformation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1129–1138.

[72] T. Huang, Y. Zhang, M. Zheng, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge diffusion for distillation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 65 299–65 316, 2023.

[73] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[74] A. M. Mansourian, A. Jalali, R. Ahmadi, and S. Kasaei, "Attention-guided feature distillation for semantic segmentation," *arXiv preprint arXiv:2403.05451*, 2024.

[75] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[76] T. Liu, C. Chen, X. Yang, and W. Tan, "Rethinking knowledge distillation with raw features for semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1155–1164.

[77] J. Yuan, M. H. Phan, L. Liu, and Y. Liu, "Fakd: Feature augmented knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 595–605.

[78] Y. Zhang, T. Huang, J. Liu, T. Jiang, K. Cheng, and S. Zhang, "Freekd: Knowledge distillation via semantic frequency prompt," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 931–15 940.

[79] X. Liu, L. Li, C. Li, and A. Yao, "Norm: Knowledge distillation via n-to-one representation matching," *arXiv preprint arXiv:2305.13803*, 2023.

[80] P. Passban, Y. Wu, M. Rezagholizadeh, and Q. Liu, "Alp-kd: Attention-based layer projection for knowledge distillation," in *Proceedings of the AAAI Conference on artificial intelligence*, vol. 35, no. 15, 2021, pp. 13 657–13 665.

[81] S. Srinivas and F. Fleuret, "Knowledge transfer with jacobian matching," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4723–4731.

[82] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 578–587.

[83] K. Xu, L. Rui, Y. Li, and L. Gu, "Feature normalized knowledge distillation for image classification," in *European conference on computer vision*. Springer, 2020, pp. 664–680.

[84] J. Xie, B. Shuai, J.-F. Hu, J. Lin, and W.-S. Zheng, "Improving fast segmentation with teacher-student learning," *arXiv preprint arXiv:1810.08476*, 2018.

[85] C. Zhang and Y. Peng, "Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification," *arXiv preprint arXiv:1804.10069*, 2018.

[86] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5007–5016.

[87] C. Wang, J. Zhong, Q. Dai, R. Li, Q. Yu, and B. Fang, "Local structure consistency and pixel-correlation distillation for compact semantic segmentation," *Applied Intelligence*, vol. 53, no. 6, pp. 6307–6323, 2023.

[88] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4794–4802.

[89] C. Wang, Z. Wang, D. Chen, S. Zhou, Y. Feng, and C. Chen, "Online adversarial knowledge distillation for graph neural networks," *Expert Systems with Applications*, vol. 237, p. 121671, 2024.

[90] Q. Guo, X. Wang, Y. Wu, Z. Yu, D. Liang, X. Hu, and P. Luo, "Online knowledge distillation via collaborative learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 020–11 029.

[91] X. Zhu, S. Gong *et al.*, "Knowledge distillation by on-the-fly native ensemble," *Advances in neural information processing systems*, vol. 31, 2018.

[92] G. Wu and S. Gong, "Peer collaborative learning for online knowledge distillation," in *Proceedings of the AAAI Conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 10 302–10 310.

[93] Z. Li, J. Ye, M. Song, Y. Huang, and Z. Pan, "Online knowledge distillation for efficient pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 740–11 750.

[94] C. Yang, Z. An, H. Zhou, F. Zhuang, Y. Xu, and Q. Zhang, "Online knowledge distillation via mutual contrastive learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 212–10 227, 2023.

[95] J. Rao, X. Meng, L. Ding, S. Qi, X. Liu, M. Zhang, and D. Tao, "Parameter-efficient and student-friendly knowledge distillation," *IEEE Transactions on Multimedia*, 2023.

[96] T. Zhang, M. Xue, J. Zhang, H. Zhang, Y. Wang, L. Cheng, J. Song, and M. Song, "Generalization matters: Loss minima flattening via parameter hybridization for efficient online knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 176–20 185.

[97] G. M. Jacob, V. Agarwal, and B. Stenger, "Online knowledge distillation for multi-task learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2359–2368.

[98] N.-A. Ypsilantis, K. Chen, A. Araujo, and O. Chum, "Udon: Universal dynamic online distillation for generic image representations," *Advances in Neural Information Processing Systems*, vol. 37, pp. 86 836–86 859, 2024.

[99] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *International conference on machine learning*. PMLR, 2018, pp. 1607–1616.

[100] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3713–3722.

[101] X. Lan, X. Zhu, and S. Gong, "Self-referenced deep learning," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*. Springer, 2019, pp. 284–300.

[102] T.-B. Xu and C.-L. Liu, "Data-distortion guided self-distillation for deep neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 5565–5572.

[103] H. Mobahi, M. Farajtabar, and P. Bartlett, "Self-distillation amplifies regularization in hilbert space," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3351–3361, 2020.

[104] L. Zhang, C. Bao, and K. Ma, "Self-distillation: Towards efficient and compact neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4388–4403, 2021.

[105] K. Kim, B. Ji, D. Yoon, and S. Hwang, "Self-knowledge distillation with progressive refinement of targets," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6567–6576.

[106] Z. Zhang and M. Sabuncu, "Self-distillation as instance-specific label smoothing," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2184–2195, 2020.

[107] Z. Yang, A. Zeng, Z. Li, T. Zhang, C. Yuan, and Y. Li, "From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 185–17 194.

[108] H. Zhang, S. Lin, W. Liu, P. Zhou, J. Tang, X. Liang, and E. P. Xing, "Iterative graph self-distillation," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[109] Z. Xiao, H. Xing, B. Zhao, R. Qu, S. Luo, P. Dai, K. Li, and Z. Zhu, "Deep contrastive representation learning with self-distillation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.

[110] Y. Zheng, C. Wang, C. Tao, S. Lin, J. Qian, and J. Wu, "Restructuring the teacher and student in self-distillation," *IEEE Transactions on Image Processing*, 2024.

[111] L. Wu, H. Lin, Z. Gao, G. Zhao, and S. Z. Li, "A teacher-free graph knowledge distillation framework with dual self-distillation," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[112] D. Chen, N. Liu, Y. Zhu, Z. Che, R. Ma, F. Zhang, X. Mou, Y. Chang, and J. Tang, "Epsd: Early pruning with self-distillation for efficient model compression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, 2024, pp. 11 258–11 266.

[113] J. Lin, L. Li, B. Yu, W. Ou, and J. Gou, "Self-distillation via intra-class compactness," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2024, pp. 139–151.

[114] Z. Li, X. Li, L. Yang, R. Song, J. Yang, and Z. Pan, "Dual teachers for self-knowledge distillation," *Pattern Recognition*, vol. 151, p. 110422, 2024.

[115] M. Liu, Y. Yu, Z. Ji, J. Han, and Z. Zhang, "Tolerant self-distillation for image classification," *Neural Networks*, vol. 174, p. 106215, 2024.

[116] P. Liang, W. Zhang, J. Wang, and Y. Guo, "Neighbor self-knowledge distillation," *Information Sciences*, vol. 654, p. 119859, 2024.

[117] Z. Qin, S. Ni, M. Zhu, Y. Jia, S. Liu, and Y. Chen, "A feature map fusion self-distillation scheme for image classification networks," *Electronics*, vol. 14, no. 1, p. 182, 2025.

[118] S. An, Q. Liao, Z. Lu, and J.-H. Xue, "Efficient semantic segmentation via self-attention and self-distillation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15 256–15 266, 2022.

[119] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[120] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[121] N. K. Bavandpour and S. Kasaei, "Class attention map distillation for efficient semantic segmentation," in *2020 International Conference on Machine Vision and Image Processing (MVIP)*. IEEE, 2020, pp. 1–6.

[122] Y. Cho and S. Kang, "Class attention transfer for semantic segmentation," in *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2022, pp. 41–45.

[123] Z. Guo, H. Yan, H. Li, and X. Lin, "Class attention transfer based knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 868–11 877.

[124] A. Karine, T. Napoléon, and M. Jridi, "Channel-spatial knowledge distillation for efficient semantic segmentation," *Pattern Recognition Letters*, vol. 180, pp. 48–54, 2024.

[125] J. Gou, L. Sun, B. Yu, S. Wan, and D. Tao, "Hierarchical multi-attention transfer for knowledge distillation," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 2, pp. 1–20, 2023.

[126] Z. Guo, P. Zhang, and P. Liang, "Sakd: Sparse attention knowledge distillation," *Image and Vision Computing*, vol. 146, p. 105020, 2024.

[127] G. Yang, S. Yu, Y. Sheng, and H. Yang, "Attention and feature transfer based knowledge distillation," *Scientific Reports*, vol. 13, no. 1, p. 18369, 2023.

[128] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[129] H. Chen, Y. Wang, C. Xu, Z. Yang, C. Liu, B. Shi, C. Xu, C. Xu, and Q. Tian, "Data-free learning of student networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3514–3522.

[130] J. Ye, Y. Ji, X. Wang, X. Gao, and M. Song, "Data-free knowledge amalgamation via group-stack dual-gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 516–12 525.

[131] G. Fang, J. Song, C. Shen, X. Wang, D. Chen, and M. Song, "Data-free adversarial distillation," *arXiv preprint arXiv:1912.11006*, 2019.

[132] Y. Choi, J. Choi, M. El-Khamy, and J. Lee, "Data-free network quantization with adversarial knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 710–711.

[133] D. Liao, X. Gao, and C. Xu, "Impartial adversarial distillation: Addressing biased data-free knowledge distillation via adaptive constrained optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 3342–3350.

[134] G. Patel, K. R. Mopuri, and Q. Qiu, "Learning to retain while acquiring: Combating distribution-shift in adversarial data-free knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7786–7794.

[135] Y. Wang, Z. Chen, D. Yang, P. Guo, K. Jiang, W. Zhang, and L. Qi, "Out of thin air: Exploring data-free adversarial robustness distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5776–5784.

[136] K. Do, T. H. Le, D. Nguyen, D. Nguyen, H. Harikumar, T. Tran, S. Rana, and S. Venkatesh, "Momentum adversarial distillation: Handling large distribution shifts in data-free knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 055–10 067, 2022.

[137] H. Zhao, X. Sun, J. Dong, M. Manic, H. Zhou, and H. Yu, "Dual discriminator adversarial distillation for data-free model compression," *International Journal of Machine Learning and Cybernetics*, pp. 1–18, 2022.

[138] Y. Zhou, Y. Zhang, L. Y. Zhang, and Z. Hua, "Derd: data-free adversarial robustness distillation through self-adversarial teacher group," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10 055–10 064.

[139] R. Liu, N. Fusi, and L. Mackey, "Teacher-student compression with generative adversarial networks," *arXiv preprint arXiv:1812.02271*, 2018.

[140] P. Micaelli and A. J. Storkey, "Zero-shot knowledge transfer via adversarial belief matching," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[141] M. Zhang, N. U. Naresh, and Y. He, "Adversarial data augmentation for task-specific knowledge distillation of pre-trained transformers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 685–11 693.

[142] Z. Xu, Y.-C. Hsu, and J. Huang, "Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks," *arXiv preprint arXiv:1709.00513*, 2017.

[143] J. Liu, Y. Chen, and K. Liu, "Exploiting the ground-truth: An adversarial imitation based knowledge distillation approach for event detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6754–6761.

[144] M. Ghorbani, F. Fooladgar, and S. Kasaei, "Be your own best competitor! multi-branched adversarial knowledge transfer," *arXiv preprint arXiv:2010.04516*, 2020.

[145] F.-A. Croitoru, N.-C. Ristea, D. Dăscălescu, R. T. Ionescu, F. S. Khan, and M. Shah, "Lightning fast video anomaly detection via multi-scale adversarial distillation," *Computer Vision and Image Understanding*, vol. 247, p. 104074, 2024.

[146] P. Liu, W. Liu, H. Ma, Z. Jiang, and M. Seok, "Ktan: knowledge transfer adversarial network," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[147] W. Wang, W. Hong, F. Wang, and J. Yu, "Gan-knowledge distillation for one-stage object detection," *IEEE Access*, vol. 8, pp. 60 719–60 727, 2020.

[148] I. Chung, S. Park, J. Kim, and N. Kwak, "Feature-map-level online adversarial knowledge distillation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2006–2015.

[149] X. Zhang, S. Lu, H. Gong, Z. Luo, and M. Liu, "Amln: adversarial-based mutual learning network for online knowledge distillation," in *European Conference on Computer Vision*. Springer, 2020, pp. 158–173.

[150] P. Li, C. Shu, Y. Xie, Y. Qu, and H. Kong, "Hierarchical knowledge squeezed adversarial network compression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 370–11 377.

[151] Z. Shen, Z. He, and X. Xue, "Meal: Multi-model ensemble via adversarial learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 4886–4893.

[152] V. Belagiannis, A. Farshad, and F. Galasso, "Adversarial network compression," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[153] Y. Wang, C. Xu, C. Xu, and D. Tao, "Adversarial learning of portable student networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[154] X. Wang, R. Zhang, Y. Sun, and J. Qi, "Adversarial distillation for learning with privileged provisions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 786–797, 2019.

[155] ——, "Kdgan: Knowledge distillation with generative adversarial networks," *Advances in neural information processing systems*, vol. 31, 2018.

[156] A. Aguinaldo, P.-Y. Chiang, A. Gain, A. Patil, K. Pearson, and S. Feizi, "Compressing gans using knowledge distillation," *arXiv preprint arXiv:1902.00159*, 2019.

[157] Y. Wang, A. Gonzalez-Garcia, D. Berga, L. Herranz, F. S. Khan, and J. v. d. Weijer, "Minegan: effective knowledge transfer from gans to target domains with few images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9332–9341.

[158] M. Li, J. Lin, Y. Ding, Z. Liu, J.-Y. Zhu, and S. Han, "Gan compression: Efficient architectures for interactive conditional gans," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5284–5294.

[159] L. Zhang, X. Chen, X. Tu, P. Wan, N. Xu, and K. Ma, "Wavelet knowledge distillation: Towards efficient image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 464–12 474.

[160] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," in *European Conference on Computer Vision*. Springer, 2024, pp. 87–103.

[161] B. Liu, P. Wang, and S. Ge, "Learning differentially private diffusion models via stochastic adversarial distillation," in *European Conference on Computer Vision*. Springer, 2024, pp. 55–71.

[162] F. Kong, J. Duan, L. Sun, H. Cheng, R. Xu, H. Shen, X. Zhu, X. Shi, and K. Xu, "Act-diffusion: Efficient adversarial consistency training for one-step diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8890–8899.

[163] S. Lin, A. Wang, and X. Yang, "Sdxl-lightning: Progressive adversarial diffusion distillation," *arXiv preprint arXiv:2402.13929*, 2024.

[164] M. Wei, J. Zhou, J. Sun, and X. Zhang, "Adversarial score distillation: When score distillation meets gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8131–8141.

[165] Z. Wan, D. Paschalidou, I. Huang, H. Liu, B. Shen, X. Xiang, J. Liao, and L. Guibas, "Cad: Photorealistic 3d generation via adversarial distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 194–10 207.

[166] M. Goldblum, L. Fowl, S. Feizi, and T. Goldstein, "Adversarially robust distillation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 3996–4003.

[167] B. Zi, S. Zhao, X. Ma, and Y.-G. Jiang, "Revisiting adversarial robustness distillation: Robust soft labels make student better," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 443–16 452.

[168] J. Zhu, J. Yao, B. Han, J. Zhang, T. Liu, G. Niu, J. Zhou, J. Xu, and H. Yang, "Reliable adversarial distillation with unreliable teachers," *arXiv preprint arXiv:2106.04928*, 2021.

[169] S. Zhao, J. Yu, Z. Sun, B. Zhang, and X. Wei, "Enhanced accuracy and robustness via multi-teacher adversarial distillation," in *European Conference on Computer Vision*. Springer, 2022, pp. 585–602.

[170] H. Lee, S. Cho, and C. Kim, "Indirect gradient matching for adversarial robust distillation," *arXiv preprint arXiv:2312.03286*, 2023.

[171] B. Huang, M. Chen, Y. Wang, J. Lu, M. Cheng, and W. Wang, "Boosting accuracy and robustness of student models via adaptive adversarial distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 668–24 677.

[172] S. Yin, Z. Xiao, M. Song, and J. Long, "Adversarial distillation based on slack matching and attribution region alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 605–24 614.

[173] H. Wang, Y. Deng, S. Yoo, H. Ling, and Y. Lin, "Agkd-bml: Defense against adversarial attack by attention guided knowledge distillation and bi-directional metric learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7658–7667.

[174] J. Dong, P. Koniusz, J. Chen, Z. J. Wang, and Y.-S. Ong, "Robust distillation via untargeted and targeted intermediate adversarial samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 432–28 442.

[175] J. Jung, H. Jang, H. Song, and J. Lee, "Peeraid: Improving adversarial distillation from a specialized peer tutor," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 482–24 491.

[176] H. Park and D. Min, "Dynamic guidance adversarial distillation with enhanced teacher knowledge," in *European Conference on Computer Vision*. Springer, 2024, pp. 204–219.

[177] J. Dong, P. Koniusz, J. Chen, and Y.-S. Ong, "Adversarially robust distillation by reducing the student-teacher variance gap," in *European Conference on Computer Vision*. Springer, 2024, pp. 92–111.

[178] T. Nguyen-Duc, T. Le, H. Zhao, J. Cai, and D. Phung, "Adversarial local distribution regularization for knowledge distillation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4681–4690.

[179] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers." in *Interspeech*, 2017, pp. 3697–3701.

[180] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *arXiv preprint arXiv:1610.05755*, 2016.

[181] C. Wang, J. Zhong, Q. Dai, Q. Yu, Y. Qi, B. Fang, and X. Li, "Mted: multiple teachers ensemble distillation for compact semantic segmentation," *Neural Computing and Applications*, vol. 35, no. 16, pp. 11 789–11 806, 2023.

[182] H. Zhang, D. Chen, and C. Wang, "Adaptive multi-teacher knowledge distillation with meta-learning," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 1943–1948.

[183] R. Shang, W. Li, S. Zhu, L. Jiao, and Y. Li, "Multi-teacher knowledge distillation based on joint guidance of probe and adaptive corrector," *Neural Networks*, vol. 164, pp. 345–356, 2023.

[184] Y.-e. Lin, S. Yin, Y. Ding, and X. Liang, "Atmkd: adaptive temperature guided multi-teacher knowledge distillation," *Multimedia Systems*, vol. 30, no. 5, p. 292, 2024.

[185] X. Cheng, Z. Zhang, W. Weng, W. Yu, and J. Zhou, "De-mkd: Decoupled multi-teacher knowledge distillation based on entropy," *Mathematics*, vol. 12, no. 11, p. 1672, 2024.

[186] Y. Wang and Y. Xu, "Relation-based multi-teacher knowledge distillation," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–6.

[187] S. Luo, X. Wang, G. Fang, Y. Hu, D. Tao, and M. Song, "Knowledge amalgamation from heterogeneous networks by common feature learning," *arXiv preprint arXiv:1906.10546*, 2019.

[188] X. Chen, J. Su, and J. Zhang, "A two-teacher framework for knowledge distillation," in *Advances in Neural Networks–ISNN 2019: 16th International Symposium on Neural Networks, ISNN 2019, Moscow, Russia, July 10–12, 2019, Proceedings, Part I 16*. Springer, 2019, pp. 58–66.

[189] S. Park and N. Kwak, "Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks," in *ECAI 2020*. IOS Press, 2020, pp. 1411–1418.

[190] U. Asif, J. Tang, and S. Harrer, "Ensemble knowledge distillation for learning improved and efficient networks," in *ECAI 2020*. IOS Press, 2020, pp. 953–960.

[191] S. Cao, M. Li, J. Hays, D. Ramanan, Y.-X. Wang, and L. Gui, "Learning lightweight object detectors via multi-teacher progressive distillation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 3577–3598.

[192] C. Shen, X. Wang, J. Song, L. Sun, and M. Song, "Amalgamating knowledge towards comprehensive classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3068–3075.

[193] C. Shen, M. Xue, X. Wang, J. Song, L. Sun, and M. Song, "Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3504–3513.

[194] J. Vongkulbhisal, P. Vinayavekhin, and M. Visentini-Scarzanella, "Unifying heterogeneous classifiers with distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3175–3184.

[195] S. Luo, W. Pan, X. Wang, D. Wang, H. Tang, and M. Song, "Collaboration by competition: Self-coordinated knowledge amalgamation for multi-talent student learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 631–646.

[196] A. Amirkhani, A. Khosravian, M. Masih-Tehrani, and H. Kashiani, "Robust semantic segmentation with multi-teacher knowledge distillation," *IEEE Access*, vol. 9, pp. 119 049–119 066, 2021.

[197] Q. Xu, L. Liu, and B. Ji, "Knowledge distillation guided by multiple homogeneous teachers," *Information Sciences*, vol. 607, pp. 230–243, 2022.

[198] H. Zhang, F. Mao, M. Xue, G. Fang, Z. Feng, J. Song, and M. Song, "Knowledge amalgamation for object detection with transformers," *IEEE Transactions on Image Processing*, vol. 32, pp. 2093–2106, 2023.

[199] L. Li, N. Jiang, J. Tang, and X. Huang, "Amalgamating knowledge for comprehensive classification with uncertainty suppression," in *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2024, pp. 1–5.

[200] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2827–2836.

[201] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 103–118.

[202] ——, "Learning with privileged information via adversarial discriminative modality distillation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2581–2593, 2019.

[203] F. Huo, W. Xu, J. Guo, H. Wang, and S. Guo, "C2kd: Bridging the modality gap for cross-modal knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 006–16 015.

[204] F. M. Thoker and J. Gall, "Cross-modal knowledge distillation for action recognition," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 6–10.

[205] L. Wang, Y. Chae, S.-H. Yoon, T.-K. Kim, and K.-J. Yoon, "Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 608–619.

[206] Z. Liu, X. Qi, and C.-W. Fu, "3d-to-2d distillation for indoor scene parsing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4464–4474.

[207] Y. Hong, H. Dai, and Y. Ding, "Cross-modality knowledge distillation network for monocular 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 87–104.

[208] S. Zhou, W. Liu, C. Hu, S. Zhou, and C. Ma, "Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5116–5125.

[209] Z. Wang, D. Li, C. Luo, C. Xie, and X. Yang, "Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8637–8646.

[210] M. Klingner, S. Borse, V. R. Kumar, B. Rezaei, V. Narayanan, S. Yogamani, and F. Porikli, "X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 343–13 353.

[211] L. Zhao, J. Song, and K. A. Skinner, "Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 470–15 480.

[212] H. Zhang, X. Yan, D. Bai, J. Gao, P. Wang, B. Liu, S. Cui, and Z. Li, "Radocc: Learning cross-modality occupancy knowledge through rendering assisted distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7060–7068.

[213] T. Zhang, H. Guo, Q. Jiao, Q. Zhang, and J. Han, "Efficient rgb-t tracking via cross-modality distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5404–5413.

[214] P. Lee, T. Kim, M. Shim, D. Wee, and H. Byun, "Decomposed cross-modal distillation for rgb-based temporal action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2373–2383.

[215] A. Andonian, S. Chen, and R. Hamid, "Robust cross-modal representation learning with progressive self-distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 430–16 441.

[216] W. Wang, X. He, Y. Zhang, L. Guo, J. Shen, J. Li, and J. Liu, "Cm-masksd: Cross-modality masked self-distillation for referring image segmentation," *IEEE Transactions on Multimedia*, 2024.

[217] P. Sarkar and A. Etemad, "Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 13, 2024, pp. 14 875–14 885.

[218] S. Lee and B. C. Song, "Graph-based knowledge distillation by multi-head attention network," *arXiv preprint arXiv:1907.02226*, 2019.

[219] N. Passalis, M. Tzelepi, and A. Tefas, "Heterogeneous knowledge distillation using information flow modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2339–2348.

[220] S. Zhou, Y. Wang, D. Chen, J. Chen, X. Wang, C. Wang, and J. Bu, "Distilling holistic knowledge with graph neural networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 387–10 396.

[221] Y. Chen, P. Chen, S. Liu, L. Wang, and J. Jia, "Deep structured instance graph for distilling object detectors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4359–4368.

[222] M. Ghorbani, M. Bahrami, A. Kazi, M. Soleymani Baghshah, H. R. Rabiee, and N. Navab, "Gkd: Semi-supervised graph knowledge distillation for graph-independent inference," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24.* Springer, 2021, pp. 709–718.

[223] S. Lee and B. C. Song, "Interpretable embedding procedure knowledge transfer via stacked principal component analysis and graph neural network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 8297–8305.

[224] Z. Luo, J.-T. Hsieh, L. Jiang, J. C. Niebles, and L. Fei-Fei, "Graph distillation for action detection with privileged modalities," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 166–183.

[225] S. Minami, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Knowledge transfer graph for deep collaborative learning," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[226] Y. Tian, S. Pei, X. Zhang, C. Zhang, and N. Chawla, "Knowledge distillation on graphs: A survey," *ACM Computing Surveys*, 2023.

[227] W. Zhang, X. Miao, Y. Shao, J. Jiang, L. Chen, O. Ruas, and B. Cui, "Reliable data distillation on graph convolutional network," in *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, 2020, pp. 1399–1414.

[228] W. Zhang, Y. Jiang, Y. Li, Z. Sheng, Y. Shen, X. Miao, L. Wang, Z. Yang, and B. Cui, "Rod: reception-aware online distillation for sparse graphs," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2232–2242.

[229] Y. Chen, Y. Bian, X. Xiao, Y. Rong, T. Xu, and J. Huang, "On self-distilling graph neural network," *arXiv preprint arXiv:2011.02255*, 2020.

[230] S. Rezayi, H. Zhao, S. Kim, R. A. Rossi, N. Lipka, and S. Li, "Edge: Enriching knowledge graph embeddings with external text," *arXiv preprint arXiv:2104.04909*, 2021.

[231] Z. Guo, C. Zhang, Y. Fan, Y. Tian, C. Zhang, and N. V. Chawla, "Boosting graph neural networks via adaptive knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, 2023, pp. 7793–7801.

[232] C. Huo, D. Jin, Y. Li, D. He, Y.-B. Yang, and L. Wu, "T2-gnn: Graph neural networks for graphs with incomplete features and structure via teacher-student distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 4339–4346.

[233] Y. Zhu, J. Li, L. Chen, and Z. Zheng, "The devil is in the data: Learning fair graph neural networks via partial knowledge distillation," in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 1012–1021.

[234] L. Yu, S. Pei, L. Ding, J. Zhou, L. Li, C. Zhang, and X. Zhang, "Sail: Self-augmented graph contrastive learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8927–8935.

[235] J. Guo, D. Chen, and C. Wang, "Alignahead: online cross-layer knowledge extraction on graph neural networks," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[236] Y. Dong, B. Zhang, Y. Yuan, N. Zou, Q. Wang, and J. Li, "Reliant: Fair knowledge distillation for graph neural networks," in *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 2023, pp. 154–162.

[237] C. Yang, Q. Wu, and J. Yan, "Geometric knowledge distillation: Topology compression for graph neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 761–29 775, 2022.

[238] H. He, J. Wang, Z. Zhang, and F. Wu, "Compressing deep graph neural networks via adversarial knowledge distillation," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 534–544.

[239] X. Deng and Z. Zhang, "Graph-free knowledge distillation for graph neural networks," *arXiv preprint arXiv:2105.07519*, 2021.

[240] C. Yang, J. Liu, and C. Shi, "Extract the knowledge of graph neural networks and go beyond it: An effective knowledge distillation framework," in *Proceedings of the web conference 2021*, 2021, pp. 1227–1237.

[241] S. Zhang, Y. Liu, Y. Sun, and N. Shah, "Graph-less neural networks: Teaching old mlps new tricks via distillation," *arXiv preprint arXiv:2110.08727*, 2021.

[242] W. Zheng, E. W. Huang, N. Rao, S. Katariya, Z. Wang, and K. Subbian, "Cold brew: Distilling graph node representations with incomplete or missing neighborhoods," *arXiv preprint arXiv:2111.04840*, 2021.

[243] Y. Tian, C. Zhang, Z. Guo, X. Zhang, and N. V. Chawla, "Nosmog: Learning noise-robust and structure-aware mlps on graphs," *arXiv preprint arXiv:2208.10010*, 2022.

[244] S. Park and Y. S. Heo, "Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy," *Sensors*, vol. 20, no. 16, p. 4616, 2020.

[245] D. Ma, K. Zhang, Q. Cao, J. Li, and X. Gao, "Coordinate attention guided dual-teacher adaptive knowledge distillation for image classification," *Expert Systems with Applications*, vol. 250, p. 123892, 2024.

[246] Z. Zhou, C. Zhuge, X. Guan, and W. Liu, "Channel distillation: Channel-wise attention for knowledge distillation," *arXiv preprint arXiv:2006.01683*, 2020.

[247] I. Sarridis, C. Koutlis, G. Kordopatis-Zilos, I. Kompatsiaris, and S. Papadopoulos, "Indistill: Information flow-preserving knowledge distillation for model compression," *arXiv preprint arXiv:2205.10003*, 2022.

[248] B. Lee, K. Ko, J. Hong, and H. Ko, "Prune channel and distill: Discriminative knowledge distillation for semantic segmentation," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 339–345.

[249] Y. Guo, W. Zhang, J. Wang, M. Ji, C. Zhen, and Z. Guo, "Afmpm: adaptive feature map pruning method based on feature distillation," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 2, pp. 573–588, 2024.

[250] T. Huang, Y. Zhang, S. You, F. Wang, C. Qian, J. Cao, and C. Xu, "Masked distillation with receptive tokens," *arXiv preprint arXiv:2205.14589*, 2022.

[251] Z. Tian, P. Chen, X. Lai, L. Jiang, S. Liu, H. Zhao, B. Yu, M.-C. Yang, and J. Jia, "Adaptive perspective distillation for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1372–1387, 2022.

[252] W. Sun, D. Chen, C. Wang, D. Ye, Y. Feng, and C. Chen, "Holistic weighted distillation for semantic segmentation," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 396–401.

[253] Y. Sun, L. Huang, Q. Zhu, and D. Liang, "Cs-kd: Confused sample knowledge distillation for semantic segmentation of aerial imagery," in *International Conference on Intelligent Computing*. Springer, 2024, pp. 266–278.

[254] S. Kim, G. Ham, S. Lee, D. Jang, and D. Kim, "Maximizing discrimination capability of knowledge distillation with energy function," *Knowledge-Based Systems*, vol. 296, p. 111911, 2024.

[255] J. Gou, X. Zhou, L. Du, Y. Zhan, W. Chen, and Z. Yi, "Difference-aware distillation for semantic segmentation," *IEEE Transactions on Multimedia*, 2024.

[256] M. Kang, S. Son, and D. Kim, "Adaptive class token knowledge distillation for efficient vision transformer," *Knowledge-Based Systems*, vol. 304, p. 112531, 2024.

[257] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[258] C. Gao, X. Wu, P. Wang, J. Wang, L. Zang, Z. Wang, and S. Hu, "Distilcse: Effective knowledge distillation for contrastive sentence embeddings," *arXiv preprint arXiv:2112.05638*, 2021.

[259] J. Fan, C. Li, X. Liu, M. Song, and A. Yao, "Augmentation-free dense contrastive knowledge distillation for efficient semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 51 359–51 370, 2023.

[260] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[261] L. Chen, D. Wang, Z. Gan, J. Liu, R. Henao, and L. Carin, "Wasserstein contrastive representation distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 296–16 305.

[262] C. Liu, D. Zhang, and K. Qin, "Knowledge distillation for single image super-resolution via contrastive learning," in *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 2024, pp. 1079–1083.

[263] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable

visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[264] Y. Chen, X. Qiao, Z. Sun, and X. Li, "Comkd-clip: Comprehensive knowledge distillation for contrastive language-image pre-traning model," *arXiv preprint arXiv:2408.04145*, 2024.

[265] X. Xing, Y. Hou, H. Li, Y. Yuan, H. Li, and M. Q.-H. Meng, "Categorical relation-preserving contrastive knowledge distillation for medical image classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer, 2021, pp. 163–173.

[266] L. Wang, J. Huang, Y. Li, K. Xu, Z. Yang, and D. Yu, "Improving weakly supervised visual grounding by contrastive knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 090–14 100.

[267] J. Zhu, S. Tang, D. Chen, S. Yu, Y. Liu, M. Rong, A. Yang, and X. Wang, "Complementary relation contrastive distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9260–9269.

[268] Y. Wei, Z. Wei, Y. Rao, J. Li, J. Zhou, and J. Lu, "Lidar distillation: Bridging the beam-induced domain gap for 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 179–195.

[269] J. Yang, S. Shi, R. Ding, Z. Wang, and X. Qi, "Towards efficient 3d object detection with knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 21 300–21 313, 2022.

[270] J. Zhang and J. Liu, "Voxel-to-pillar: Knowledge distillation of 3d object detection in point cloud," in *Proceedings of the 4th European Symposium on Software Engineering*, 2023, pp. 99–104.

[271] Y. Li, S. Xu, M. Lin, J. Yin, B. Zhang, and X. Cao, "Representation disparity-aware distillation for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6715–6724.

[272] L. Zhang, R. Dong, H.-S. Tai, and K. Ma, "Pointdistiller: structured knowledge distillation towards efficient and compact 3d detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21 791–21 801.

[273] H. Cho, J. Choi, G. Baek, and W. Hwang, "itkd: Interchange transfer-based knowledge distillation for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 540–13 549.

[274] A. Gambashidze, A. Dadukin, M. Golyadkin, M. Razzhivina, and I. Makarov, "Weak-to-strong 3d object detection with x-ray distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 055–15 064.

[275] G. Bang, K. Choi, J. Kim, D. Kum, and J. W. Choi, "Radardistill: Boosting radar-based object detection performance via knowledge distillation from lidar features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 491–15 500.

[276] X. Huang, H. Wu, X. Li, X. Fan, C. Wen, and C. Wang, "Sunshine to rainstorm: Cross-weather knowledge distillation for robust 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2409–2416.

[277] J. Zeng, L. Chen, H. Deng, L. Lu, J. Yan, Y. Qiao, and H. Li, "Distilling focal knowledge from imperfect expert for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 992–1001.

[278] Y. Lee and W. Wu, "Feature adversarial distillation for point cloud classification," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 970–974.

[279] K. Zhou, M. Dong, P. Zhi, and S. Wang, "Cascaded network with hierarchical self-distillation for sparse point cloud classification," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[280] Q. Zheng, C. Zhang, and J. Sun, "Efficient point cloud classification via offline distillation framework and negative-weight self-distillation technique," *arXiv preprint arXiv:2409.02020*, 2024.

[281] Z. Tian, W. Li, J. Hu, and C. Deng, "Joint graph entropy knowledge distillation for point cloud classification and robustness against corruptions," *Information Sciences*, vol. 648, p. 119542, 2023.

[282] S. Qiu, F. Jiang, H. Zhang, X. Xue, and J. Pu, "Multi-to-single knowledge distillation for point cloud semantic segmentation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9303–9309.

[283] J. Cen, S. Zhang, Y. Pei, K. Li, H. Zheng, M. Luo, Y. Zhang, and Q. Chen, "Cmdfusion: Bidirectional fusion network with cross-modality knowledge distillation for lidar semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 771–778, 2023.

[284] F. Jiang, H. Gao, S. Qiu, H. Zhang, R. Wan, and J. Pu, "Knowledge distillation from 3d to bird's-eye-view for lidar semantic segmentation," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 402–407.

[285] A. Adamyan and E. Harutyunyan, "Smaller3d: Smaller models for 3d semantic segmentation using minkowski engine and knowledge distillation methods," *arXiv preprint arXiv:2305.03188*, 2023.

[286] Y. Liu, L. Kong, J. Cen, R. Chen, W. Zhang, L. Pan, K. Chen, and Z. Liu, "Segment any point cloud sequences by distilling vision foundation models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[287] Y. Hou, X. Zhu, Y. Ma, C. C. Loy, and Y. Li, "Point-to-voxel knowledge distillation for lidar semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8479–8488.

[288] K. Genova, X. Yin, A. Kundu, C. Pantofaru, F. Cole, A. Sud, B. Brewington, B. Shucker, and T. Funkhouser, "Learning 3d semantic segmentation with only 2d image supervision," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 361–372.

[289] Y. Zhang, Y. Qu, Y. Xie, Z. Li, S. Zheng, and C. Li, "Perturbed self-distillation: Weakly supervised large-scale point cloud semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 520–15 528.

[290] Z. Yang, R. Li, E. Ling, C. Zhang, Y. Wang, D. Huang, K. T. Ma, M. Hur, and G. Lin, "Label-guided knowledge distillation for continual semantic segmentation on 2d images and 3d point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 601–18 612.

[291] A. Umam, C.-K. Yang, M.-H. Chen, J.-H. Chuang, and Y.-Y. Lin, "Partdistill: 3d shape part segmentation by vision-language model distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3470–3479.

[292] J. Chen, J. Wang, Y. Shi, N. Ling, and B. Yin, "Mvp-net: Multi-view depth image guided cross-modal distillation network for point cloud upsampling," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 9759–9768.

[293] H. Sun, N. Vysotskaya, T. Sukianto, H. Feng, J. Ott, X. Peng, L. Servadei, and R. Wille, "Lircdepth: Lightweight radar-camera depth estimation via knowledge distillation and uncertainty guidance," *arXiv preprint arXiv:2412.16380*, 2024.

[294] J. Xiao, K. Zhang, X. Xu, S. Liu, S. Wu, Z. Huang, and L. Li, "Improving accuracy and efficiency of monocular depth estimation in power grid environments using point cloud optimization and knowledge distillation," *Energies*, vol. 17, no. 16, p. 4068, 2024.

[295] Y.-C. Liu, Y.-K. Huang, H.-Y. Chiang, H.-T. Su, Z.-Y. Liu, C.-T. Chen, C.-Y. Tseng, and W. H. Hsu, "Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining," *arXiv preprint arXiv:2104.04687*, 2021.

[296] K. Fu, P. Gao, R. Zhang, H. Li, Y. Qiao, and M. Wang, "Distillation with contrast is all you need for self-supervised point cloud representation learning," *arXiv preprint arXiv:2202.04241*, 2022.

[297] C. Sautier, G. Puy, S. Gidaris, A. Boulch, A. Bursuc, and R. Marlet, "Image-to-lidar self-supervised distillation for autonomous driving data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9891–9901.

[298] X. Yan, H. Zhan, C. Zheng, J. Gao, R. Zhang, S. Cui, and Z. Li, "Let images give you more: Point cloud cross-modal training for shape analysis," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 398–32 411, 2022.

[299] Z. Qi, R. Dong, G. Fan, Z. Ge, X. Zhang, K. Ma, and L. Yi, "Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 223–28 243.

[300] S. Yan, C. Song, Y. Kong, and Q. Huang, "Multi-view representation is what you need for point-cloud pre-training," *arXiv preprint arXiv:2306.02558*, 2023.

[301] S. Wang, W. Li, W. Liu, X. Liu, and J. Zhu, "Lidar2map: In defense of lidar-based semantic map construction using online camera distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5186–5195.

[302] Z. Zhang, Y. Dong, Y. Liu, and L. Yi, "Complete-to-partial 4d distillation for self-supervised point cloud sequence representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 661–17 670.

[303] S. Zhang, J. Deng, L. Bai, H. Li, W. Ouyang, and Y. Zhang, "Hvdistill: Transferring knowledge from images to point clouds via unsupervised hybrid-view distillation," *International Journal of Computer Vision*, pp. 1–15, 2024.

[304] Y. Yao, Y. Zhang, Z. Yin, J. Luo, W. Ouyang, and X. Huang, "3d point cloud pre-training with knowledge distilled from 2d images," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[305] R. Huang, X. Pan, H. Zheng, H. Jiang, Z. Xie, C. Wu, S. Song, and G. Huang, "Joint representation learning for text and 3d point cloud," *Pattern Recognition*, vol. 147, p. 110086, 2024.

[306] N. S. Dutt, S. Muralikrishnan, and N. J. Mitra, "Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4494–4504.

[307] P.-C. Yu, C. Sun, and M. Sun, "Data efficient 3d learner via knowledge transferred from 2d model," in *European Conference on Computer Vision*. Springer, 2022, pp. 182–198.

[308] R. Zhang, L. Wang, Y. Qiao, P. Gao, and H. Li, "Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 769–21 780.

[309] A. Cardace, R. Spezialetti, P. Z. Ramirez, S. Salti, and L. Di Stefano, "Self-distillation for unsupervised 3d domain adaptation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4166–4177.

[310] Y. Wu, M. Xing, Y. Zhang, Y. Xie, J. Fan, Z. Shi, and Y. Qu, "Cross-modal unsupervised domain adaptation for 3d semantic segmentation via bidirectional fusion-then-distillation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 490–498.

[311] Q. Li, X. Peng, C. Yan, P. Gao, and Q. Hao, "Self-ensembling for 3d point cloud domain adaptation," *Image and Vision Computing*, vol. 154, p. 105409, 2025.

[312] S. Wang, R. She, Q. Kang, X. Jian, K. Zhao, Y. Song, and W. P. Tay, "Distilvpr: Cross-modal knowledge distillation for visual place recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 9, 2024, pp. 10 377–10 385.

[313] Q. Zhang, J. Hou, and Y. Qian, "Pointmcd: Boosting deep point cloud encoders via multi-view cross-modal distillation for 3d shape recognition," *IEEE Transactions on Multimedia*, 2023.

[314] Z. Xia, Y. Liu, X. Li, X. Zhu, Y. Ma, Y. Li, Y. Hou, and Y. Qiao, "Scpnet: Semantic scene completion on point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 642–17 651.

[315] Z. Fan, Y. He, Z. Wang, K. Wu, H. Liu, and J. He, "Reconstruction-aware prior distillation for semi-supervised point cloud completion," *arXiv preprint arXiv:2204.09186*, 2022.

[316] F. Lin, H. Liu, H. Zhou, S. Hou, K. D. Yamada, G. S. Fischer, Y. Li, H. K. Zhang, and Z. Zhang, "Loss distillation via gradient matching for point cloud completion with weighted chamfer distance," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 511–518.

[317] L. Wang, D. Lin, K. Yang, R. Liu, Q. Guo, W. Xie, M. Wang, L. Liang, Y. Wang, and P. Li, "Voxel proposal network via multi-frame knowledge distillation for semantic scene completion," *Advances in Neural Information Processing Systems*, vol. 37, pp. 101 096–101 115, 2025.

[318] J. Huang, H. Zheng, and X. Feng, "Multi-scale distillation for low scanline resolution depth completion," in *2024 9th International Conference on Computer and Communication Systems (ICCCS)*. IEEE, 2024, pp. 854–859.

[319] J. Kim, J. Noh, M. Jeong, W. Lee, Y. Park, and J. Park, "Adnet: Non-local affinity distillation network for lightweight depth completion with guidance from missing lidar points," *IEEE Robotics and Automation Letters*, 2024.

[320] S. Hwang, J. Lee, W. J. Kim, S. Woo, K. Lee, and S. Lee, "Lidar depth completion using color-embedded information via knowledge distillation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 482–14 496, 2021.

[321] T. Y. Liu, P. Agrawal, A. Chen, B.-W. Hong, and A. Wong, "Monitored distillation for positive congruent depth completion," in *European Conference on Computer Vision*. Springer, 2022, pp. 35–53.

[322] K. Zhou, Z. Tan, S. Yang, and S. Wang, "Enhancing the encoding process in point cloud completion," in *Proceedings of the 2024 13th International Conference on Computing and Pattern Recognition*, 2024, pp. 59–65.

[323] Q. Zhang, X. Liu, W. Li, H. Chen, J. Liu, J. Hu, Z. Xiong, C. Yuan, and Y. Wang, "Distilling semantic priors from sam to efficient image restoration models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 409–25 419.

[324] S. Wang, J. Yu, W. Li, W. Liu, X. Liu, J. Chen, and J. Zhu, "Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 792–14 801.

[325] L. Jiang, Y. Li, Y. Liu, Z. Dong, M. Yao, and Y. Lin, "Knowledge distillation-based point cloud registration method," in *International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2023)*, vol. 13105. SPIE, 2024, pp. 648–655.

[326] C. Löwens, T. Funke, A. Wagner, and A. P. Condurache, "Unsupervised point cloud registration with self-distillation," *arXiv preprint arXiv:2409.07558*, 2024.

[327] X. Yi, Z. Wu, Q. Xu, P. Zhou, J.-H. Lim, and H. Zhang, "Diffusion time-step curriculum for one image to 3d generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9948–9958.

[328] D. Decatur, I. Lang, K. Aberman, and R. Hanocka, "3d paintbrush: Local stylization of 3d shapes with cascaded score distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4473–4483.

[329] T. Van Vo, M. N. Vu, B. Huang, T. Nguyen, N. Le, T. Vo, and A. Nguyen, "Open-vocabulary affordance detection using knowledge distillation and text-point correlation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 968–13 975.

[330] T. Huang, J. Zhang, J. Chen, Y. Liu, and Y. Liu, "Resolution-free point cloud sampling network with data distillation," in *European Conference on Computer Vision*. Springer, 2022, pp. 54–70.

[331] Y. Ding, Q. Zhu, X. Liu, W. Yuan, H. Zhang, and C. Zhang, "Kd-mvs: Knowledge distillation based self-supervised learning for multi-view stereo," in *European conference on computer vision*. Springer, 2022, pp. 630–646.

[332] Y. Tian, S. Sun, and J. Tang, "Multi-view teacher–student network," *Neural Networks*, vol. 146, pp. 69–84, 2022.

[333] J. Li, M. Lu, J. Liu, Y. Guo, L. Du, and S. Zhang, "Bev-lgkd: A unified lidar-guided knowledge distillation framework for bev 3d object detection," *arXiv preprint arXiv:2212.00623*, 2022.

[334] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Bevdistill: Cross-modal bev distillation for multi-view 3d object detection," *arXiv preprint arXiv:2211.09386*, 2022.

[335] S. Jang, D. U. Jo, S. J. Hwang, D. Lee, and D. Ji, "Stxd: structural and temporal cross-modal distillation for multi-view 3d object detection," *Advances in Neural Information Processing Systems*, vol. 36, pp. 29 323–29 342, 2023.

[336] C. Acar, K. Binici, A. Tekirdağ, and Y. Wu, "Visual-policy learning through multi-camera view to single-camera view knowledge distillation for robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 691–698, 2023.

[337] L. Zhang, Y. Shi, K. Wang, Z. Zhang, H.-S. Tai, Y. He, and K. Ma, "Structured knowledge distillation towards efficient multi-view 3d object detection." in *BMVC*, 2023, pp. 339–344.

[338] C. Wang, J. Zhong, Q. Dai, Y. Qi, F. Shi, B. Fang, and X. Li, "Multi-view knowledge distillation for efficient semantic segmentation," *Journal of Real-Time Image Processing*, vol. 20, no. 2, p. 39, 2023.

[339] Y.-C. Lin and V. S. Tseng, "Multi-view knowledge distillation transformer for human action recognition," *arXiv preprint arXiv:2303.14358*, 2023.

[340] Z. Jiang, J. Zhang, Y. Zhang, Q. Liu, Z. Hu, B. Wang, and Y. Wang, "Fsd-bev: Foreground self-distillation for multi-view 3d object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 110–126.

[341] H. Zhao, Q. Zhang, S. Zhao, Z. Chen, J. Zhang, and D. Tao, "Simdistill: Simulated multi-modal distillation for bev 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7460–7468.

[342] L. Xu, Q. Cui, R. Hong, W. Xu, E. Chen, X. Yuan, C. Li, and Y. Tang, "Group multi-view transformer for 3d shape analysis with spatial encoding," *IEEE Transactions on Multimedia*, 2024.

[343] Y. Qiang, X. Dong, X. Liu, and Y. Yang, "Mt-mv-kdf: A novel multi-task multi-view knowledge distillation framework for myocardial infarction detection and localization," *Biomedical Signal Processing and Control*, vol. 95, p. 106382, 2024.

[344] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," *arXiv preprint arXiv:1606.07947*, 2016.

[345] X. Tan, Y. Ren, D. He, T. Qin, Z. Zhao, and T.-Y. Liu, "Multilingual neural machine translation with knowledge distillation," *arXiv preprint arXiv:1902.10461*, 2019.

[346] F. Wang, J. Yan, F. Meng, and J. Zhou, "Selective knowledge distillation for neural machine translation," *arXiv preprint arXiv:2105.12967*, 2021.

[347] X. Liang, L. Wu, J. Li, T. Qin, M. Zhang, and T.-Y. Liu, "Multi-teacher distillation with single model for neural machine translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 992–1002, 2022.

[348] H. Jin, S. Son, J. Park, Y. Kim, H. Noh, and Y. Lee, "Align-to-distill: Trainable attention alignment for knowledge distillation in neural machine translation," *arXiv preprint arXiv:2403.01479*, 2024.

[349] M. Hu, Y. Peng, F. Wei, Z. Huang, D. Li, N. Yang, and M. Zhou, "Attention-guided answer distillation for machine reading comprehension," *arXiv preprint arXiv:1808.07644*, 2018.

[350] G. Izacard and E. Grave, "Distilling knowledge from reader to retriever for question answering," *arXiv preprint arXiv:2012.04584*, 2020.

[351] Z. Yang, L. Shou, M. Gong, W. Lin, and D. Jiang, "Model compression with two-stage multi-teacher knowledge distillation for web question answering system," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 690–698.

[352] J. Liu, Y. Chen, and J. Xu, "Machine reading comprehension as data augmentation: A case study on implicit event argument extraction," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 2716–2725.

[353] L. Yin, L. Wang, Z. Cai, S. Lu, R. Wang, A. AlSanad, S. A. AlQahtani, X. Chen, Z. Yin, X. Li *et al.*, "Dpal-bert: A faster and lighter question answering model." *CMES-Computer Modeling in Engineering & Sciences*, vol. 141, no. 1, 2024.

[354] Y.-C. Chen, Z. Gan, Y. Cheng, J. Liu, and J. Liu, "Distilling knowledge learned in bert for text generation," *arXiv preprint arXiv:1911.03829*, 2019.

[355] M. A. Haidar and M. Rezagholizadeh, "Textkd-gan: Text generation using knowledge distillation and generative adversarial networks," in *Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019, Kingston, ON, Canada, May 28–31, 2019, Proceedings 32*. Springer, 2019, pp. 107–118.

[356] J. Ren, H. Peng, L. Jiang, J. Wu, Y. Tong, L. Wang, X. Bai, B. Wang, and Q. Yang, "Transferring knowledge distillation for multilingual social event detection," *arXiv preprint arXiv:2108.03084*, 2021.

[357] P. Yu, H. Ji, and P. Natarajan, "Lifelong event detection with knowledge transfer," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 5278–5290.

[358] S. Shakeri, A. Sethy, and C. Cheng, "Knowledge distillation in document retrieval," *arXiv preprint arXiv:1911.11065*, 2019.

[359] X. Chen, B. He, K. Hui, L. Sun, and Y. Sun, "Simplified tinybert: Knowledge distillation for document retrieval," in *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer, 2021, pp. 241–248.

[360] A. K. Bhunia, A. Sain, P. N. Chowdhury, and Y.-Z. Song, "Text is text, no matter what: Unifying text recognition using knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 983–992.

[361] Z.-R. Wang and J. Du, "Joint architecture and knowledge distillation in cnn for chinese text recognition," *Pattern Recognition*, vol. 111, p. 107722, 2021.

[362] S. Liang, M. Gong, J. Pei, L. Shou, W. Zuo, X. Zuo, and D. Jiang, "Reinforced iterative knowledge distillation for cross-lingual named entity recognition," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 3231–3239.

[363] L. Ge, C. Hu, G. Ma, J. Liu, and H. Zhang, "Discrepancy and uncertainty aware denoising knowledge distillation for zero-shot cross-lingual named entity recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 18 056–18 064.

[364] K.-Y. Chen, S.-H. Liu, B. Chen, and H.-M. Wang, "An information distillation framework for extractive summarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 161–170, 2017.

[365] Y. Liu, S. Shen, and M. Lapata, "Noisy self-knowledge distillation for text summarization," *arXiv preprint arXiv:2009.07032*, 2020.

[366] T. T. Nguyen and A. T. Luu, "Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 103–11 111.

[367] X. Liu, P. He, W. Chen, and J. Gao, "Improving multi-task deep neural networks via knowledge distillation for natural language understanding," *arXiv preprint arXiv:1904.09482*, 2019.

[368] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from bert into simple neural networks," *arXiv preprint arXiv:1903.12136*, 2019.

[369] J. FitzGerald, S. Ananthakrishnan, K. Arkoudas, D. Bernardi, A. Bhagia, C. Delli Bovi, J. Cao, R. Chada, A. Chauhan, L. Chen *et al.*, "Alexa teacher model: Pretraining and distilling multi-billion-parameter encoders for natural language understanding systems," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2893–2902.

[370] J. Lehečka, J. Švec, P. Ircing, and L. Šmídl, "Bert-based sentiment analysis using distillation," in *International conference on statistical language and speech processing*. Springer, 2020, pp. 58–70.

[371] G. Malki, "Efficient sentiment analysis and topic modeling in nlp using knowledge distillation and transfer learning," 2023.

[372] R. Xu and Y. Yang, "Cross-lingual distillation for text classification," *arXiv preprint arXiv:1705.02073*, 2017.

[373] Y. Li and W. Li, "Data distillation for text classification," *arXiv preprint arXiv:2104.08448*, 2021.

[374] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition." in *Interspeech*, 2016, pp. 3439–3443.

[375] K. Markov and T. Matsui, "Robust speech recognition using generalized distillation framework." in *Interspeech*, 2016, pp. 2364–2368.

[376] R. Takashima, S. Li, and H. Kawai, "An investigation of a knowledge distillation method for ctc acoustic models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5809–5813. [Online]. Available: https://ieeexplore.ieee.org/document/8461995/

[377] R. M. Mun'im, N. Inoue, and K. Shinoda, "Sequence-level knowledge distillation for model compression of attention-based sequence-to-sequence speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6151–6155.

[378] G. Kurata and G. Saon, "Knowledge distillation from offline to streaming rnn transducer for end-to-end speech recognition." in *Interspeech*, 2020, pp. 2117–2121.

[379] J. W. Yoon, H. Lee, H. Y. Kim, W. I. Cho, and N. S. Kim, "Tutornet: Towards flexible knowledge distillation for end-to-end speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1626–1638, 2021.

[380] C. You, N. Chen, and Y. Zou, "Knowledge distillation for improved accuracy in spoken question answering," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7793–7797.

[381] J. W. Yoon, B. J. Woo, S. Ahn, H. Lee, and N. S. Kim, "Inter-kd: Intermediate knowledge distillation for ctc-based automatic speech recognition," in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, 2022, pp. 1–8. [Online]. Available: https://arxiv.org/abs/2211.15075

[382] K. Zhao, H. D. Nguyen, A. Jain, N. Susanj, A. Mouchtaris, L. Gupta, and M. Zhao, "Knowledge distillation via module replacing for automatic speech recognition with recurrent neural network transducer," in *23rd Interspeech Conference*, 2022.

[383] J.-U. Kim, S.-H. Kim, H.-Y. Kim, H.-J. Kim, and H.-G. Kang, "Knowledge distillation-based training of speech enhancement for noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 149–160, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10535505/

[384] J.-S. Park, J.-H. Cho, H.-G. Kim, and J.-H. Kim, "End-to-end emotional speech recognition using acoustic model adaptation based on knowledge distillation," *Multimedia Tools and Applications*, vol. 82, pp. 14 681–14 697, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s11042-023-14680-y

[385] Y. Zhang, L. Liu, and L. Liu, "Cuing without sharing: A federated cued speech recognition framework via mutual knowledge distillation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8781–8789.

[386] S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, "Student-teacher network learning with enhanced features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5275–5279.

[387] J. Wu, Y. Hua, S. Yang, H. Qin, and H. Qin, "Speech enhancement using generative adversarial network by distilling knowledge from statistical method," *Applied Sciences*, vol. 9, no. 16, p. 3396, 2019.

[388] S. Kim and M. Kim, "Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 176–180.

[389] W. Shin, B. H. Lee, J. S. Kim, H. J. Park, and S. W. Han, "Metricgan-okd: multi-metric optimization of metricgan via online knowledge distillation for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2023, pp. 31 521–31 538.

[390] H. J. Park, W. Shin, J. S. Kim, and S. W. Han, "Leveraging non-causal knowledge via cross-network knowledge distillation for real-time speech enhancement," *IEEE Signal Processing Letters*, 2024.

[391] V. Mingote, A. Miguel, D. Ribas, A. Ortega, and E. Lleida, "Knowledge distillation and random erasing data augmentation for text-dependent speaker verification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6824–6828.

[392] Z. Peng, X. He, K. Ding, T. Lee, and G. Wan, "Label-free knowledge distillation with contrastive loss for light-weight speaker recognition," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2022, pp. 324–328.

[393] B. Liu, H. Wang, Z. Chen, S. Wang, and Y. Qian, "Self-knowledge distillation via feature enhancement for speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7542–7546.

[394] Y. Jin, G. Hu, H. Chen, D. Miao, L. Hu, and C. Zhao, "Cross-modal distillation for speaker recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 977–12 985.

[395] V. Mingote, A. Miguel, A. Ortega, and E. Lleida, "Class token and knowledge distillation for multi-head self-attention speaker verification systems," *Digital Signal Processing*, vol. 133, p. 103859, 2023.

[396] D.-T. Truong, R. Tao, J. Q. Yip, K. A. Lee, and E. S. Chng, "Emphasized non-target speaker knowledge in knowledge distillation for automatic speaker verification," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 336–10 340.

[397] K. A. Hoang, K. Duong, T. N. V. Minh, T. Le, and H. T. Nguyen, "Integrating voice activity detection to enhance robustness of on-device speaker verification," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2024, pp. 369–380.

[398] Y. Liu, H. Xiong, Z. He, J. Zhang, H. Wu, H. Wang, and C. Zong, "End-to-end speech translation with knowledge distillation," *arXiv preprint arXiv:1904.08075*, 2019.

[399] M. Gaido, M. A. Di Gangi, M. Negri, and M. Turchi, "End-to-end speech-translation with knowledge distillation: Fbk@ iwslt2020," *arXiv preprint arXiv:2006.02965*, 2020.

[400] H. Inaguma, T. Kawahara, and S. Watanabe, "Source and target bidirectional knowledge distillation for end-to-end speech translation," *arXiv preprint arXiv:2104.06457*, 2021.

[401] Y. Lei, Z. Xue, X. Zhao, H. Sun, S. Zhu, X. Lin, and D. Xiong, "Ckdst: Comprehensively and effectively distill knowledge from machine translation to end-to-end speech translation," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 3123–3137.

[402] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 3918–3926.

[403] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T.-Y. Liu, "Lrspeech: Extremely low-resource speech synthesis and recognition," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2802–2812.

[404] Y. A. Li, R. Kumar, and Z. Jin, "Dmdspeech: Distilled diffusion model surpassing the teacher in zero-shot speech synthesis via direct metric optimization," *arXiv preprint arXiv:2410.11097*, 2024.

[405] X. Chen, G. Liu, J. Shi, J. Xu, and B. Xu, "Distilled binary neural network for monaural speech separation," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[406] J. Zhang, C. Zorila, R. Doddipatla, and J. Barker, "Teacher-student mixit for unsupervised and semi-supervised speech separation," *arXiv preprint arXiv:2106.07843*, 2021.

[407] P. Shen, X. Lu, S. Li, and H. Kawai, "Feature representation of short utterances based on knowledge distillation for spoken language identification." in *Interspeech*, 2018, pp. 1813–1817.

[408] ——, "Interactive learning of teacher-student model for short utterance spoken language identification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5981–5985.

[409] ——, "Knowledge distillation-based representation learning for short-utterance spoken language identification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2674–2683, 2020.

[410] S. Kim, G. Kim, S. Shin, and S. Lee, "Two-stage textual knowledge distillation for end-to-end spoken language understanding," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7463–7467.

[411] U. Cappellazzo, D. Falavigna, and A. Brutti, "An investigation of the combination of rehearsal and knowledge distillation in continual learning for spoken language understanding," *arXiv preprint arXiv:2211.08161*, 2022.

[412] T. Mao and C. Zhang, "Diffslu: Knowledge distillation based diffusion model for cross-lingual spoken language understanding," in *Proc. of Interspeech*, 2023.

[413] U. Cappellazzo, M. Yang, D. Falavigna, and A. Brutti, "Sequence-level knowledge distillation for class-incremental end-to-end spoken language understanding," *arXiv preprint arXiv:2305.13899*, 2023.

[414] Y.-L. Liao, X. Chen, C.-C. Wang, and J.-S. R. Jang, "Adversarial speaker distillation for countermeasure model on automatic speaker verification," *arXiv preprint arXiv:2203.17031*, 2022.

[415] J. Xue, C. Fan, J. Yi, C. Wang, Z. Wen, D. Zhang, and Z. Lv, "Learning from yourself: A self-distillation method for fake speech detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[416] Y. Ren, H. Peng, L. Li, and Y. Yang, "Lightweight voice spoofing detection using improved one-class learning and knowledge distillation," *IEEE Transactions on Multimedia*, 2023.

[417] Y. Ren, H. Peng, L. Li, X. Xue, Y. Lan, and Y. Yang, "A voice spoofing detection framework for iot systems with feature pyramid and online knowledge distillation," *Journal of Systems Architecture*, vol. 143, p. 102981, 2023.

[418] J. Lu, Y. Zhang, W. Wang, Z. Shang, and P. Zhang, "One-class knowledge distillation for spoofing speech detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 251–11 255.

[419] T. M. Wani and I. Amerini, "Audio deepfake detection: A continual approach with feature distillation and dynamic class rebalancing," in *International Conference on Pattern Recognition*. Springer, 2025, pp. 211–227.

[420] C.-C. Chang, C.-C. Kao, M. Sun, and C. Wang, "Intra-utterance similarity preserving knowledge distillation for audio tagging," *arXiv preprint arXiv:2009.01759*, 2020. [Online]. Available: https://arxiv.org/abs/2009.01759

[421] Y. Yin, H. Shrivastava, Y. Zhang, Z. Liu, R. R. Shah, and R. Zimmermann, "Enhanced audio tagging via multi-to single-modal teacher-student mutual learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 10 709–10 717.

[422] F. Schmid, K. Koutini, and G. Widmer, "Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation," in *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[423] Y. Liang, Y. Long, Y. Li, J. Liang, and Y. Wang, "Joint framework with deep feature distillation and adaptive focal loss for weakly supervised audio tagging and acoustic event detection," *Digital Signal Processing*, vol. 123, p. 103446, 2022.

[424] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "Ced: Consistent ensemble distillation for audio tagging," *arXiv preprint arXiv:2308.11957*, 2023. [Online]. Available: https://arxiv.org/abs/2308.11957

[425] Y. Tang, Z. Ma, and H. Zhang, "Enhanced feature learning with normalized knowledge distillation for audio tagging," in *Proc. Interspeech 2024*, 2024, pp. 1695–1699.

[426] C. You, N. Chen, F. Liu, D. Yang, and Y. Zou, "Towards data distillation for end-to-end spoken conversational question answering," *arXiv preprint arXiv:2010.08923*, 2020.

[427] C. You, N. Chen, and Y. Zou, "Contextualized attention-based knowledge transfer for spoken conversational question answering," *arXiv preprint arXiv:2010.11066*, 2020.

[428] ——, "Mrd-net: Multi-modal residual knowledge distillation for spoken question answering." in *IJCAI*, 2021, pp. 3985–3991.

[429] X. Xu, H. Liu, M. Wu, W. Wang, and M. D. Plumbley, "Efficient audio captioning with encoder-level knowledge distillation," *arXiv preprint arXiv:2407.14329*, 2024.

[430] P. Primus and G. Widmer, "A knowledge distillation approach to improving language-based audio retrieval models," DCASE2024 Challenge, Tech. Rep, Tech. Rep., 2024.

[431] G. Wang, P. Zhao, Y. Shi, C. Zhao, and S. Yang, "Generative model-based feature knowledge distillation for action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 14, 2024, pp. 15 474–15 482.

[432] M. Liu, X. Chen, Y. Zhang, Y. Li, and J. M. Rehg, "Attention distillation for learning video representations," *arXiv preprint arXiv:1904.03249*, 2019.

[433] Y. Jiang, Z. Zhang, J. Wei, C.-M. Feng, G. Li, X. Wan, S. Cui, and Z. Li, "Let video teaches you more: Video-to-image knowledge distillation using detection transformer for medical video lesion detection," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 944–949.

[434] X. Li, S. He, J. Wu, Y. Yu, L. Nie, and M. Zhang, "Mask again: Masked knowledge distillation for masked video modeling," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2221–2232.

[435] C. Wang and Z. Tang, "The staged knowledge distillation in video classification: Harmonizing student progress by a complementary weakly supervised framework," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[436] S. Bhardwaj, M. Srinivasan, and M. M. Khapra, "Efficient video classification using fewer frames," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 354–363.

[437] H. Kim, S. Lee, H. Kang, and S. Im, "Offline-to-online knowledge distillation for video instance segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 159–168.

[438] J. Dong, M. Zhang, Z. Zhang, X. Chen, D. Liu, X. Qu, X. Wang, and B. Liu, "Dual learning with dynamic knowledge distillation for partially relevant video retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 302–11 312.

[439] F. Camarena, M. Gonzalez-Mendoza, and L. Chang, "Knowledge distillation in video-based human action recognition: An intuitive approach to efficient and flexible model training," *Journal of Imaging*, vol. 10, no. 4, p. 85, 2024.

[440] M.-C. Wu and C.-T. Chiu, "Multi-teacher knowledge distillation for compressed video action recognition based on deep learning," *Journal of systems architecture*, vol. 103, p. 101695, 2020.

[441] E. Soufleri, D. Ravikumar, and K. Roy, "Advancing compressed video action recognition through progressive knowledge distillation," *arXiv preprint arXiv:2407.02713*, 2024.

[442] R. Miles, M. K. Yucel, B. Manganelli, and A. Saa-Garriga, "Mobilevos: Real-time video object segmentation contrastive learning meets knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 480–10 490.

[443] A. Monti, A. Porrello, S. Calderara, P. Coscia, L. Ballan, and R. Cucchiara, "How many observations are enough? knowledge distillation for trajectory forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6553–6562.

[444] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, L. Yuan, and Y.-G. Jiang, "Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6312–6322.

[445] C. Feng, D. Danier, H. Wang, F. Zhang, B. Vallade, A. Mackin, and D. Bull, "Rankdvqa-mini: knowledge distillation-driven deep video quality assessment," in *2024 Picture Coding Symposium (PCS)*. IEEE, 2024, pp. 1–5.

[446] Y. Wang, D. Zeng, S. Wada, and S. Kurihara, "Videoadviser: Video knowledge distillation for multimodal transfer learning," *IEEE Access*, vol. 11, pp. 51 229–51 240, 2023.

[447] T. Afouras, J. S. Chung, and A. Zisserman, "Asr is all you need: Cross-modal distillation for lip reading," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2143–2147.

[448] A. Perez, V. Sanguineti, P. Morerio, and V. Murino, "Audio-visual model distillation using acoustic images," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 2854–2863.

[449] K. Fu, P. Shi, Y. Song, S. Ge, X. Lu, and J. Li, "Ultrafast video attention prediction with coupled knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 802–10 809.

[450] R. T. Mullapudi, S. Chen, K. Zhang, D. Ramanan, and K. Fatahalian, "Online model distillation for efficient video inference," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2019, pp. 3573–3582.

[451] B. Pan, H. Cai, D.-A. Huang, K.-H. Lee, A. Gaidon, E. Adeli, and J. C. Niebles, "Spatio-temporal graph for video captioning with knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 870–10 879.

[452] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," *arXiv preprint arXiv:1909.10351*, 2019.

[453] V. Sanh, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[454] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for bert model compression," *arXiv preprint arXiv:1908.09355*, 2019.

[455] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: On the importance of pre-training compact models," *arXiv preprint arXiv:1908.08962*, 2019.

[456] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "Mobilebert: a compact task-agnostic bert for resource-limited devices," *arXiv preprint arXiv:2004.02984*, 2020.

[457] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in Neural Information Processing Systems*, vol. 33, 5776–5788, 2020.

[458] S. Mukherjee and A. Awadallah, "Xtremedistil: Multi-stage distillation for massive multilingual models," *arXiv preprint arXiv:2004.05686*, 2020.

[459] S. Mukherjee, A. H. Awadallah, and J. Gao, "Xtremedistiltransformers: Task transfer for task-agnostic distillation," *arXiv preprint arXiv:2106.04563*, 2021.

[460] W. Zhou, C. Xu, and J. McAuley, "Bert learns to teach: Knowledge distillation with meta learning," *arXiv preprint arXiv:2106.04570*, 2021.

[461] C. Liang, S. Zuo, Q. Zhang, P. He, W. Chen, and T. Zhao, "Less is more: Task-aware layer-wise distillation for language model compression," in *International Conference on Machine Learning*. PMLR, 2023, pp. 20 852–20 867.

[462] R. Agarwal, N. Vieillard, P. Stanczyk, S. Ramos, M. Geist, and O. Bachem, "Gkd: Generalized knowledge distillation for auto-regressive sequence models," *arXiv preprint arXiv:2306.13649*, 2023.

[463] R. Agarwal, N. Vieillard, Y. Zhou, P. Stanczyk, S. R. Garea, M. Geist, and O. Bachem, "On-policy distillation of language models: Learning from self-generated mistakes," in *The Twelfth International Conference on Learning Representations*, 2024.

[464] I. Hwang, H. Park, Y. Lee, J. Yang, and S. Maeng, "Pc-lora: Low-rank adaptation for progressive model compression with knowledge distillation," *arXiv preprint arXiv:2406.09117*, 2024.

[465] Y. Gu, L. Dong, F. Wei, and M. Huang, "Minillm: Knowledge distillation of large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[466] S. Li, J. Chen, Y. Shen, Z. Chen, X. Zhang, Z. Li, H. Wang, J. Qian, B. Peng, Y. Mao *et al.*, "Explanations from large language models make small reasoners better," *arXiv preprint arXiv:2210.06726*, 2022.

[467] L. C. Magister, J. Mallinson, J. Adamek, E. Malmi, and A. Severyn, "Teaching small language models to reason," *arXiv preprint arXiv:2212.08410*, 2022.

[468] N. Ho, L. Schmid, and S.-Y. Yun, "Large language models are reasoning teachers," *arXiv preprint arXiv:2212.10071*, 2022.

[469] H. Chen, S. Wu, X. Quan, R. Wang, M. Yan, and J. Zhang, "Mcc-kd: Multi-cot consistent knowledge distillation," *arXiv preprint arXiv:2310.14747*, 2023.

[470] C.-Y. Hsieh, C.-L. Li, C.-K. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, and T. Pfister, "Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes," *arXiv preprint arXiv:2305.02301*, 2023.

[471] L. H. Li, J. Hessel, Y. Yu, X. Ren, K.-W. Chang, and Y. Choi, "Symbolic chain-of-thought distillation: Small models can also" think" step-by-step," *arXiv preprint arXiv:2306.14050*, 2023.

[472] P. Wang, Z. Wang, Z. Li, Y. Gao, B. Yin, and X. Ren, "Scott: Self-consistent chain-of-thought distillation," *arXiv preprint arXiv:2305.01879*, 2023.

[473] H. Chae, Y. Song, K. T.-i. Ong, T. Kwon, M. Kim, Y. Yu, D. Lee, D. Kang, and J. Yeo, "Dialogue chain-of-thought distillation for commonsense-aware conversational agents," *arXiv preprint arXiv:2310.09343*, 2023.

[474] X. Zhu, B. Qi, K. Zhang, X. Long, and B. Zhou, "Pad: Program-aided distillation specializes large models in reasoning," *arXiv preprint arXiv:2305.13888*, 2023.

[475] Y. Li, P. Yuan, S. Feng, B. Pan, B. Sun, X. Wang, H. Wang, and K. Li, "Turning dust into gold: Distilling complex reasoning capabilities from llms by leveraging negative data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 18 591–18 599.

[476] J. C.-Y. Chen, Z. Wang, H. Palangi, R. Han, S. Ebrahimi, L. Le, V. Perot, S. Mishra, M. Bansal, C.-Y. Lee *et al.*, "Reverse thinking makes llms stronger reasoners," *arXiv preprint arXiv:2411.19865*, 2024.

[477] X. Zhuang, Z. Zhu, Z. Wang, X. Cheng, and Y. Zou, "Unicott: A unified framework for structural chain-of-thought distillation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[478] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.

[479] W. Zhou, S. Zhang, Y. Gu, M. Chen, and H. Poon, "Universalner: Targeted distillation from large language models for open named entity recognition," *arXiv preprint arXiv:2308.03279*, 2023.

[480] M. Wu, A. Waheed, C. Zhang, M. Abdul-Mageed, and A. F. Aji, "Lamini-lm: A diverse herd of distilled models from large-scale instructions," *arXiv preprint arXiv:2304.14402*, 2023.

[481] Y. Jiang, C. Chan, M. Chen, and W. Wang, "Lion: Adversarial distillation of proprietary large language models," *arXiv preprint arXiv:2305.12870*, 2023.

[482] M. Li, L. Chen, J. Chen, S. He, J. Gu, and T. Zhou, "Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 16 189–16 211.

[483] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[484] Y. Huang, Y. Chen, Z. Yu, and K. McKeown, "In-context learning distillation: Transferring few-shot learning ability of pre-trained language models," *arXiv preprint arXiv:2212.10670*, 2022.

[485] L. Wang, N. Yang, and F. Wei, "Learning to retrieve in-context examples for large language models," *arXiv preprint arXiv:2307.07164*, 2023.

[486] Y. Liu, "Learning to reason with autoregressive in-context distillation," in *The Second Tiny Papers Track at ICLR 2024*, 2024.

[487] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[488] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[489] A. Radford, "Improving language understanding by generative pre-training," 2018.

[490] J. A. Baktash and M. Dawodi, "Gpt-4: A review on advancements and opportunities in natural language processing," *arXiv preprint arXiv:2305.03195*, 2023.

[491] M. D. M. Reddy, M. S. M. Basha, M. M. C. Hari, and M. N. Penchalaiah, "Dall-e: Creating images from text," *UGC Care Group I Journal*, vol. 8, no. 14, pp. 71–75, 2021.

[492] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[493] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[494] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[495] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[496] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[497] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[498] Z. Chen, W. Wang, Z. Zhao, F. Su, A. Men, and H. Meng, "Practicaldg: Perturbation distillation on vision-language models for hybrid domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 501–23 511.

[499] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen *et al.*, "Simple open-vocabulary object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 728–755.

[500] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[501] Z. Ma, G. Luo, J. Gao, L. Li, Y. Chen, S. Wang, C. Zhang, and W. Hu, "Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 074–14 083.

[502] H. Bangalath, M. Maaz, M. U. Khattak, S. H. Khan, and F. Shahbaz Khan, "Bridging the gap between object and image-level representations for open-vocabulary detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 781–33 794, 2022.

[503] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Open-vocabulary detr with conditional matching," in *European Conference on Computer Vision*. Springer, 2022, pp. 106–122.

[504] J. Xie and S. Zheng, "Zero-shot object detection through vision-language embedding alignment," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2022, pp. 1–15.

[505] D. Kim, A. Angelova, and W. Kuo, "Region-aware pretraining for open-vocabulary object detection with vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 144–11 154.

[506] S. Wu, W. Zhang, S. Jin, W. Liu, and C. C. Loy, "Aligning bag of regions for open-vocabulary object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 254–15 264.

[507] L. Wang, Y. Liu, P. Du, Z. Ding, Y. Liao, Q. Qi, B. Chen, and S. Liu, "Object-aware distillation pyramid for open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 186–11 196.

[508] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 084–14 093.

[509] C. Feng, Y. Zhong, Z. Jie, X. Chu, H. Ren, X. Wei, W. Xie, and L. Ma, "Promptdet: Towards open-vocabulary detection using uncurated images," in *European Conference on Computer Vision*. Springer, 2022, pp. 701–717.

[510] M. Gao, C. Xing, J. C. Niebles, J. Li, R. Xu, W. Liu, and C. Xiong, "Open vocabulary object detection with pseudo bounding-box labels," in *European Conference on Computer Vision*. Springer, 2022, pp. 266–282.

[511] D. Huynh, J. Kuen, Z. Lin, J. Gu, and E. Elhamifar, "Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7020–7031.

[512] Y. Long, J. Han, R. Huang, H. Xu, Y. Zhu, C. Xu, and X. Liang, "Fine-grained visual–text prompt-driven self-training for open-vocabulary object detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[513] S. Zhao, Z. Zhang, S. Schulter, L. Zhao, B. Vijay Kumar, A. Stathopoulos, M. Chandraker, and D. N. Metaxas, "Exploiting unlabeled data with vision and language models for object detection," in *European conference on computer vision*. Springer, 2022, pp. 159–175.

[514] C. Lin, P. Sun, Y. Jiang, P. Luo, L. Qu, G. Haffari, Z. Yuan, and J. Cai, "Learning object-language alignments for open-vocabulary object detection," *arXiv preprint arXiv:2211.14843*, 2022.

[515] W. Kuo, Y. Cui, X. Gu, A. Piergiovanni, and A. Angelova, "F-vlm: Open-vocabulary object detection upon frozen vision and language models," *arXiv preprint arXiv:2209.15639*, 2022.

[516] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *European Conference on Computer Vision*. Springer, 2022, pp. 350–368.

[517] L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. Xu, and H. Xu, "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9125–9138, 2022.

[518] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7086–7096.

[519] J. Ding, N. Xue, G.-S. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 583–11 592.

[520] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," *arXiv preprint arXiv:2201.03546*, 2022.

[521] N. Zabari and Y. Hoshen, "Semantic segmentation in-the-wild without seeing any segmentation examples," *arXiv preprint arXiv:2112.03185*, 2021.

[522] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *European Conference on Computer Vision*. Springer, 2022, pp. 696–712.

[523] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu, "Zegclip: Towards adapting clip for zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 175–11 185.

[524] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7061–7070.

[525] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *European Conference on Computer Vision*. Springer, 2022, pp. 736–753.

[526] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *European Conference on Computer Vision*. Springer, 2022, pp. 540–557.

[527] G. Shin, W. Xie, and S. Albanie, "Reco: Retrieve and co-segment for zero-shot transfer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 754–33 767, 2022.

[528] J. Qin, J. Wu, P. Yan, M. Li, R. Yuxi, X. Xiao, Y. Wang, R. Wang, S. Wen, X. Pan *et al.*, "Freeseg: Unified, universal and open-vocabulary image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 446–19 455.

[529] M. Wysoczańska, M. Ramamonjisoa, T. Trzciński, and O. Siméoni, "Clip-diy: Clip dense inference yields open-vocabulary semantic segmentation for-free," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1403–1413.

[530] J. Chen, D. Zhu, G. Qian, B. Ghanem, Z. Yan, C. Zhu, F. Xiao, S. C. Culatana, and M. Elhoseiny, "Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 699–710.

[531] J. Choi, S. Lee, S. Lee, M. Lee, and H. Shim, "Understanding multi-granularity for open-vocabulary part segmentation," *arXiv preprint arXiv:2406.11384*, 2024.

[532] Y. Wei, Z. Hu, L. Shen, Z. Wang, C. Yuan, and D. Tao, "Open-vocabulary customization from clip via data-free knowledge distillation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[533] J. Xie, X. Hou, K. Ye, and L. Shen, "Clims: Cross language image matching for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4483–4492.

[534] Y. Lin, M. Chen, W. Wang, B. Wu, K. Li, B. Lin, H. Liu, and X. He, "Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 305–15 314.

[535] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Learning multi-modal class-specific tokens for weakly supervised dense object localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 596–19 605.

[536] S. Deng, W. Zhuo, J. Xie, and L. Shen, "Qa-clims: Question-answer cross language image matching for weakly supervised semantic segmentation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5572–5583.

[537] B. Murugesan, R. Hussain, R. Bhattacharya, I. Ben Ayed, and J. Dolz, "Prompting classes: exploring the power of prompt class learning in weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 291–302.

[538] B. Zhang, S. Yu, Y. Wei, Y. Zhao, and J. Xiao, "Frozen clip: A strong backbone for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3796–3806.

[539] L. Zhu, X. Wang, J. Feng, T. Cheng, Y. Li, B. Jiang, D. Zhang, and J. Han, "Weakclip: Adapting clip for weakly-supervised semantic segmentation," *International Journal of Computer Vision*, pp. 1–21, 2024.

[540] S. Jang, J. Yun, J. Kwon, E. Lee, and Y. Kim, "Dial: Dense image-text alignment for weakly supervised semantic segmentation," *arXiv preprint arXiv:2409.15801*, 2024.

[541] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[542] ——, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[543] Z. Li, X. Li, X. Fu, X. Zhang, W. Wang, S. Chen, and J. Yang, "Promptkd: Unsupervised prompt distillation for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 617–26 626.

[544] C. Ge, R. Huang, M. Xie, Z. Lai, S. Song, S. Li, and G. Huang, "Domain adaptation via prompt learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[545] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 082–18 091.

[546] R. Abdal, P. Zhu, J. Femiani, N. Mitra, and P. Wonka, "Clip2stylegan: Unsupervised extraction of stylegan edit directions," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–9.

[547] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–13, 2022.

[548] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2085–2094.

[549] J. Hyung, S. Hwang, D. Kim, H. Lee, and J. Choo, "Local 3d editing via 3d distillation of clip knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 674–12 684.

[550] Z. Canfes, M. F. Atasoy, A. Dirik, and P. Yanardag, "Text and image guided 3d avatar generation and manipulation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4421–4431.

[551] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, and Z. Liu, "Avatarclip: Zero-shot text-driven generation and animation of 3d avatars," *arXiv preprint arXiv:2205.08535*, 2022.

[552] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, "Zero-shot text-guided object generation with dream fields," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 867–876.

[553] S. Kobayashi, E. Matsumoto, and V. Sitzmann, "Decomposing nerf for editing via feature field distillation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 311–23 330, 2022.

[554] O. Michel, R. Bar-On, R. Liu, S. Benaim, and R. Hanocka, "Text2mesh: Text-driven neural stylization for meshes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 492–13 502.

[555] A. Sanghi, H. Chu, J. G. Lambourne, Y. Wang, C.-Y. Cheng, M. Fumero, and K. R. Malekshan, "Clip-forge: Towards zero-shot text-to-shape generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 603–18 613.

[556] C. Wang, M. Chai, M. He, D. Chen, and J. Liao, "Clip-nerf: Text-and-image driven manipulation of neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3835–3844.

[557] F. Liu, M. Kim, Z. Ren, and X. Liu, "Distilling clip with dual guidance for learning discriminative human body shape representation,"

in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 256–266.

[558] B. Wan and T. Tuytelaars, "Exploiting clip for zero-shot hoi detection requires knowledge distillation at multiple levels," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1805–1815.

[559] X. Li, Y. Fang, M. Liu, Z. Ling, Z. Tu, and H. Su, "Distilling large vision-language model with out-of-distribution generalizability," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2492–2503.

[560] Z. Huang, A. Zhou, Z. Ling, M. Cai, H. Wang, and Y. J. Lee, "A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 685–11 695.

[561] R. Pei, J. Liu, W. Li, B. Shao, S. Xu, P. Dai, J. Lu, and Y. Yan, "Clipping: Distilling clip-based models with a student base for video-language retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 983–18 992.

[562] C. Cuttano, G. Rosi, G. Trivigno, and G. Averta, "What does clip know about peeling a banana?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2238–2247.

[563] D. Han, S. Seo, E. Park, S.-U. Nam, and N. Kwak, "Unleash the potential of clip for video highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8275–8279.

[564] E. Son and S. J. Lee, "Cabins: Clip-based adaptive bins for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4557–4567.

[565] C. Yang, Z. An, L. Huang, J. Bi, X. Yu, H. Yang, and Y. Xu, "Clip-kd: An empirical study of distilling clip models," *arXiv preprint arXiv:2307.12732*, 2023.

[566] L. Nair, "Clip-embed-kd: Computationally efficient knowledge distillation using embeddings as teachers," *arXiv preprint arXiv:2404.06170*, 2024.

[567] K. Wu, H. Peng, Z. Zhou, B. Xiao, M. Liu, L. Yuan, H. Xuan, M. Valenzuela, X. S. Chen, X. Wang *et al.*, "Tinyclip: Clip distillation via affinity mimicking and weight inheritance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 970–21 980.

[568] X. Dong, J. Bao, Y. Zheng, T. Zhang, D. Chen, H. Yang, M. Zeng, W. Zhang, L. Yuan, D. Chen *et al.*, "Maskclip: Masked self-distillation advances contrastive language-image pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 995–11 005.

[569] C. Hetang, H. Xue, C. Le, T. Yue, W. Wang, and Y. He, "Segment anything model for road network graph extraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2556–2566.

[570] M. M. Rahman, M. Munir, D. Jha, U. Bagci, and R. Marculescu, "Ppsam: Perturbed prompts for robust adaption of segment anything model for polyp segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4989–4995.

[571] H. Kweon and K.-J. Yoon, "From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 499–19 509.

[572] P.-T. Jiang and Y. Yang, "Segment anything is a good pseudo-label generator for weakly supervised semantic segmentation," *arXiv preprint arXiv:2305.01275*, 2023.

[573] Z. Chen and Q. Sun, "Weakly-supervised semantic segmentation with image-level labels: from traditional models to foundation models," *ACM Computing Surveys*, 2023.

[574] T. Chen, Z. Mai, R. Li, and W.-l. Chao, "Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation," *arXiv preprint arXiv:2305.05803*, 2023.

[575] W. Sun, Z. Liu, Y. Zhang, Y. Zhong, and N. Barnes, "An alternative to wsss? an empirical study of the segment anything model (sam) on weakly-supervised semantic segmentation problems," *arXiv preprint arXiv:2305.01586*, 2023.

[576] S. Yin and L. Jiang, "Distilling knowledge from multiple foundation models for zero-shot image classification," *PloS one*, vol. 19, no. 9, p. e0310730, 2024.

[577] M. Ranzinger, G. Heinrich, J. Kautz, and P. Molchanov, "Am-radio: Agglomerative vision foundation model reduce all domains into one," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 490–12 500.

[578] X. Sun, P. Zhang, P. Zhang, H. Shah, K. Saenko, and X. Xia, "Dime-fm: Distilling multimodal and efficient foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 521–15 533.

[579] J. Ye, N. Wang, and X. Wang, "Featurenerf: Learning generalizable nerfs by distilling foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8962–8973.

[580] J. Shang, K. Schmeckpeper, B. B. May, M. V. Minniti, T. Kelestemur, D. Watkins, and L. Herlant, "Theia: Distilling diverse vision foundation models for robot learning," *arXiv preprint arXiv:2407.20179*, 2024.

[581] T. Gao, W. Ao, X.-A. Wang, Y. Zhao, P. Ma, M. Xie, H. Fu, J. Ren, and Z. Gao, "Enrich distill and fuse: Generalized few-shot semantic segmentation in remote sensing leveraging foundation model's assistance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2771–2780.

[582] B. B. Englert, F. J. Piva, T. Kerssies, D. De Geus, and G. Dubbelman, "Exploring the benefits of vision foundation models for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1172–1180.

[583] A. Rai, K. Buettner, and A. Kovashka, "Strategies to leverage foundational model knowledge in object affordance grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1714–1723.

[584] H. Wang, P. K. A. Vasu, F. Faghri, R. Vemulapalli, M. Farajtabar, S. Mehta, M. Rastegari, O. Tuzel, and H. Pouransari, "Sam-clip: Merging vision foundation models towards semantic and spatial understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3635–3647.

[585] S. Aleem, F. Wang, M. Maniparambil, E. Arazo, J. Dietlmeier, K. Curran, N. E. Connor, and S. Little, "Test-time adaptation with salip: A cascade of sam and clip for zero-shot medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5184–5193.

[586] X. Yang and X. Gong, "Foundation model assisted weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 523–532.

[587] M. Wysoczańska, O. Siméoni, M. Ramamonjisoa, A. Bursuc, T. Trzciński, and P. Pérez, "Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 320–337.

[588] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 729–19 739.

[589] Q. Jiang, F. Li, Z. Zeng, T. Ren, S. Liu, and L. Zhang, "T-rex2: Towards generic object detection via text-visual prompt synergy," in *European Conference on Computer Vision*. Springer, 2024, pp. 38–57.

[590] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[591] A. M. Das, R. Chaudhry, K. Kundu, and D. Modolo, "Prompting foundational models for omni-supervised instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1583–1592.

[592] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[593] H. Lin, G. Han, J. Ma, S. Huang, X. Lin, and S.-F. Chang, "Supervised masked knowledge distillation for few-shot transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 649–19 659.

[594] S. Ren, F. Wei, Z. Zhang, and H. Hu, "Tinymim: An empirical study of distilling mim pre-trained models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3687–3697.

[595] C. Shang, H. Li, F. Meng, Q. Wu, H. Qiu, and L. Wang, "Incrementer: Transformer for class-incremental semantic segmentation with knowledge distillation focusing on old class," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7214–7224.

[596] D. Kang, P. Koniusz, M. Cho, and N. Murray, "Distilling self-supervised vision transformers for weakly-supervised few-shot classification & segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 627–19 638.

[597] Y. Bai, Z. Wang, J. Xiao, C. Wei, H. Wang, A. L. Yuille, Y. Zhou, and C. Xie, "Masked autoencoders enable efficient knowledge distillers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 24 256–24 265.

[598] B. Zhao, R. Song, and J. Liang, "Cumulative spatial knowledge distillation for vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6146–6155.

[599] Z. Yang, Z. Li, A. Zeng, Z. Li, C. Yuan, and Y. Li, "Vitkd: Feature-based knowledge distillation for vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1379–1388.

[600] S. Kara, H. Ammar, J. Denize, F. Chabot, and Q.-C. Pham, "Diod: Self-distillation meets object discovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3975–3985.

[601] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce, "Localizing objects with self-supervised transformers and no labels," *arXiv preprint arXiv:2109.14279*, 2021.

[602] O. Siméoni, C. Sekkat, G. Puy, A. Vobeckỳ, É. Zablocki, and P. Pérez, "Unsupervised object localization: Observing the background to discover objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3176–3186.

[603] X. Wang, R. Girdhar, S. X. Yu, and I. Misra, "Cut and learn for unsupervised object detection and instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 3124–3134.

[604] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 38–55.

[605] N. Amini-Naieni, T. Han, and A. Zisserman, "Countgd: Multi-modal open-world counting," *Advances in Neural Information Processing Systems*, vol. 37, pp. 48 810–48 837, 2024.

[606] L. Karazija, I. Laina, A. Vedaldi, and C. Rupprecht, "Diffusion models for zero-shot open-vocabulary segmentation," *arXiv preprint arXiv:2306.09316*, 2023.

[607] J. Chang, S. Wang, H.-M. Xu, Z. Chen, C. Yang, and F. Zhao, "Detrdistill: A universal knowledge distillation framework for detr-families," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6898–6908.

[608] Y. Wang, X. Li, S. Weng, G. Zhang, H. Yue, H. Feng, J. Han, and E. Ding, "Kd-detr: Knowledge distillation for detection transformer with consistent distillation points sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 016–16 025.

[609] R. Ahmadi and S. Kasaei, "Leveraging swin transformer for local-to-global weakly supervised semantic segmentation," in *2024 13th Iranian/3rd International Machine Vision and Image Processing Conference (MVIP)*. IEEE, 2024, pp. 1–7.

[610] B. Zhang, J. Xiao, and Y. Zhao, "Dynamic feature regularized loss for weakly supervised semantic segmentation," *arXiv preprint arXiv:2108.01296*, 2021.

[611] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in neural information processing systems*, vol. 34, pp. 17 864–17 875, 2021.

[612] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4486–4497, 2021.

[613] L. Lin, H. Fan, Z. Zhang, Y. Xu, and H. Ling, "Swintrack: A simple and strong baseline for transformer tracking," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 743–16 754, 2022.

[614] W. Shi, J. Xu, and P. Gao, "Ssformer: A lightweight transformer for semantic segmentation," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2022, pp. 1–5.

[615] T. Li, H. Wang, G. Li, S. Liu, and L. Tang, "Swinf: Swin transformer with feature fusion in target detection," in *Journal of Physics: Conference Series*, vol. 2284, no. 1. IOP Publishing, 2022, p. 012027.

[616] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 730–20 740.

[617] A. Rasoulian, S. Salari, and Y. Xiao, "Weakly supervised intracranial hemorrhage segmentation using head-wise gradient-infused self-attention maps from a swin transformer in categorical learning," *arXiv preprint arXiv:2304.04902*, 2023.

[618] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 6202–6212.

[619] T. Chen and L. Mo, "Swin-fusion: swin-transformer with feature fusion for human action recognition," *Neural Processing Letters*, vol. 55, no. 8, pp. 11 109–11 130, 2023.

[620] C. Wang, T. Endo, T. Hirofuchi, and T. Ikegami, "Pyramid swin transformer: Different-size windows swin transformer for image classification and object detection." in *VISIGRAPP (5: VISAPP)*, 2023, pp. 583–590.

[621] Y. Chen, B. Zou, Z. Guo, Y. Huang, Y. Huang, F. Qin, Q. Li, and C. Wang, "Scunet++: Swin-unet and cnn bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism ct image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7759–7767.

[622] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[623] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[624] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[625] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[626] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6707–6717.

[627] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[628] H. Xu, J. Fang, X. Zhang, L. Xie, X. Wang, W. Dai, H. Xiong, and Q. Tian, "Bag of instances aggregation boosts self-supervised distillation," *arXiv preprint arXiv:2107.01691*, 2021.

[629] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[630] K. Song, J. Xie, S. Zhang, and Z. Luo, "Multi-mode online knowledge distillation for self-supervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 848–11 857.

[631] Y. Gao, J.-X. Zhuang, S. Lin, H. Cheng, X. Sun, K. Li, and C. Shen, "Disco: Remedying self-supervised learning on lightweight models with distilled contrastive learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 237–253.

[632] J. Gu, W. Liu, and Y. Tian, "Simple distillation baselines for improving small self-supervised models," *arXiv preprint arXiv:2106.11304*, 2021.

[633] S. Abbasi Koohpayegani, A. Tejankar, and H. Pirsiavash, "Compress: Self-supervised learning by compressing representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 980–12 992, 2020.

[634] A. Tejankar, S. A. Koohpayegani, V. Pillai, P. Favaro, and H. Pirsiavash, "Isd: Self-supervised learning by iterative similarity distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9609–9618.

[635] K. Navaneet, S. A. Koohpayegani, A. Tejankar, and H. Pirsiavash, "Simreg: Regression as a simple yet effective tool for self-supervised knowledge distillation," *arXiv preprint arXiv:2201.05131*, 2022.

[636] Z. Fang, J. Wang, L. Wang, L. Zhang, Y. Yang, and Z. Liu, "Seed: Self-supervised distillation for visual representation," *arXiv preprint arXiv:2101.04731*, 2021.

[637] A. Dadashzadeh, A. Whone, and M. Mirmehdi, "Auxiliary learning for self-supervised video representation via similarity-based knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4231–4240.

[638] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[639] W. Luo, "A comprehensive survey on knowledge distillation of diffusion models," *arXiv preprint arXiv:2304.04262*, 2023.

[640] Y.-T. Hsiao, S. Khodadadeh, K. Duarte, W.-A. Lin, H. Qu, M. Kwon, and R. Kalarot, "Plug-and-play diffusion distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 743–13 752.

[641] W. Feng, C. Yang, Z. An, L. Huang, B. Diao, F. Wang, and Y. Xu, "Relational diffusion distillation for efficient image generation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 205–213.

[642] D. Zhang, S. Li, C. Chen, Q. Xie, and H. Lu, "Laptop-diff: Layer pruning and normalized distillation for compressing diffusion models," *arXiv preprint arXiv:2404.11098*, 2024.

[643] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," *arXiv preprint arXiv:2202.00512*, 2022.

[644] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park, "One-step diffusion with distribution matching distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 6613–6623.

[645] K. A. Mekonnen, N. Dall'Asen, and P. Rota, "Adv-kd: Adversarial knowledge distillation for faster diffusion sampling," *arXiv preprint arXiv:2405.20675*, 2024.

[646] S. Xie, Z. Xiao, D. Kingma, T. Hou, Y. N. Wu, K. P. Murphy, T. Salimans, B. Poole, and R. Gao, "Em distillation for one-step diffusion models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 45 073–45 104, 2024.

[647] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," 2023.

[648] Y. Song and P. Dhariwal, "Improved techniques for training consistency models," *arXiv preprint arXiv:2310.14189*, 2023.

[649] Z. Geng, A. Pokle, W. Luo, J. Lin, and J. Z. Kolter, "Consistency models made easy," *arXiv preprint arXiv:2406.14548*, 2024.

[650] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.

[651] Q. Xie, Z. Liao, Z. Deng, H. Lu *et al.*, "Tlcm: Training-efficient latent consistency model for image generation with 2-8 steps," *arXiv preprint arXiv:2406.05768*, 2024.

[652] C. Lu and Y. Song, "Simplifying, stabilizing and scaling continuous-time consistency models," *arXiv preprint arXiv:2410.11081*, 2024.

[653] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8406–8441, 2023.

[654] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[655] W. Luo, T. Hu, S. Zhang, J. Sun, Z. Li, and Z. Zhang, "Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 76 525–76 546, 2023.

[656] A. López-Cifuentes, M. Escudero-Viñolo, J. Bescós, and J. C. S. Miguel, "Attention-based knowledge distillation in scene recognition: The impact of a dct-driven loss," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4769–4783, 2023.

[657] Z. Zhou and Q. Dong, "Self-distilled feature aggregation for self-supervised monocular depth estimation," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 13661. Springer, 2022, pp. 709–726.

[658] J. Hu, C. Fan, H. Jiang, X. Guo, Y. Gao, X. Lu, and T. L. Lam, "Boosting lightweight depth estimation via knowledge distillation," in *Knowledge Science, Engineering and Management*, Z. Jin, Y. Jiang, R. A. Buchmann, Y. Bi, A.-M. Ghiran, and W. Ma, Eds. Springer Nature Switzerland, 2023, pp. 27–39.

[659] Y. Xu, Y. Yang, and L. Zhang, "Multi-task learning with knowledge distillation for dense prediction," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 21 493–21 502.

[660] X. Shi, G. Dikov, G. Reitmayr, T.-K. Kim, and M. Ghafoorian, "3d distillation: Improving self-supervised monocular depth estimation on reflective surfaces," in *ICCV*. IEEE, 2023, pp. 9099–9109.

[661] W. Ka, J. Y. Lee, J. Choi, and J. Kim, "Stereo-matching knowledge distilled monocular depth estimation filtered by multiple disparity consistency," in *ICASSP*. IEEE, 2024, pp. 3840–3844.

[662] N.-H. Wang and Y.-L. Liu, "Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation," in *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[663] R. Peng, R. Wang, Y. Lai, L. Tang, and Y. Cai, "Excavating the potential capacity of self-supervised monocular depth estimation," in *ICCV*. IEEE, 2021, pp. 15 540–15 549.

[664] C. Zhao, M. Poggi, F. Tosi, L. Zhou, Q. Sun, Y. Tang, and S. Mattoccia, "Gasmono: Geometry-aided self-supervised monocular depth estimation for indoor scenes," in *ICCV*. IEEE, 2023, pp. 16 163–16 174.

[665] R. Marsal, F. Chabot, A. Loesch, W. Grolleau, and H. Sahbi, "Monoprob: Self-supervised monocular depth estimation with interpretable uncertainty," in *WACV*. IEEE, 2024, pp. 3625–3634.

[666] H. Hu, Y. Feng, D. Li, S. Zhang, and H. Zhao, "Monocular depth estimation via self-supervised self-distillation," *Sensors*, vol. 24, no. 13, p. 4090, 2024.

[667] F. Boutros, V. Struc, and N. Damer, "Adadistill: Adaptive knowledge distillation for deep face recognition," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 15113. Springer, 2024, pp. 163–182.

[668] Y.-J. Dong, F.-L. Zhang, and S.-H. Zhang, "Mal: Motion-aware loss with temporal and distillation hints for self-supervised depth estimation," in *ICRA*. IEEE, 2024, pp. 7318–7324.

[669] R. Chen, H. Luo, F. Zhao, J. Yu, Y. Jia, J. Wang, and X. Ma, "Structure-centric robust monocular depth estimation via knowledge distillation," in *ACCV*, ser. Lecture Notes in Computer Science, vol. 15480. Springer, 2024, pp. 123–140.

[670] Z. Wu, Y. Wu, J. Pu, X. Li, and X. Wang, "Attention-based depth distillation with 3d-aware positional encoding for monocular 3d object detection," in *AAAI*. AAAI Press, 2023, pp. 2892–2900.

[671] S. Shao, Z. Pei, W. Chen, R. Li, Z. Liu, and Z. Li, "Urcdc-depth: Uncertainty rectified cross-distillation with cutflip for monocular depth estimation," *IEEE Trans. Multim.*, vol. 26, pp. 3341–3353, 2024.

[672] S. Shin, J. Lee, J. Lee, Y. Yu, and K. Lee, "Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 13672. Springer, 2022, pp. 631–647.

[673] H. Otroshi-Shahreza, A. George, and S. Marcel, "Synthdistill: Face recognition with knowledge distillation from synthetic data," in *IJCB*. IEEE, 2023, pp. 1–10.

[674] P. C. Neto, I. Colakovic, S. Karakatič, and A. F. Sequeira, "Synthetic gap mitigation using knowledge distillation in fair face recognition," in *Proceedings of the ECCV 2024 Workshop on Synthetic Data for Computer Vision*, 2024.

[675] E. Caldeira, J. S. Cardoso, A. F. Sequeira, and P. Neto, "MST-KD: Multiple specialized teachers knowledge distillation for fair face recognition," in *Proceedings of the 7th Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW) at ECCV*, 2024.

[676] J. Li, Z. Guo, H. Li, S. Han, J.-W. Baek, M. Yang, R. Yang, and S. Suh, "Rethinking feature-based knowledge distillation for face recognition," in *CVPR*. IEEE, 2023, pp. 20 156–20 165.

[677] Y. Huang, Y. Wang, L. Yang, and L. Wang, "Enhanced face recognition using intra-class incoherence constraint," in *ICLR*, 2024.

[678] Z. Babnik, F. Boutros, N. Damer, P. Peer, and V. Struc, "Ai-kd: Towards alignment invariant face image quality assessment using knowledge distillation," in *IWBF*. IEEE, 2024, pp. 1–6.

[679] S. Jung, D. Lee, T. Park, and T. Moon, "Fair feature distillation for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 115–12 125.

[680] B. Liu, S. Zhang, G. Song, H. You, and Y. Liu, "Rectifying the data bias in knowledge distillation," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 1477–1486.

[681] J. Xu, S. Li, A. Deng, M. Xiong, J. Wu, J. Wu, S. Ding, and B. Hooi, "Probabilistic knowledge distillation of face ensembles," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 3489–3498.

[682] Y. Huang, J. Wu, X. Xu, and S. Ding, "Evaluation-oriented knowledge distillation for deep face recognition," in *CVPR*. IEEE, 2022, pp. 18 719–18 728.

[683] X. Di, Y. Zheng, X. Liu, and Y. Cheng, "Pros: Facial omni-representation learning via prototype-based self-distillation," in *WACV*. IEEE, 2024, pp. 6075–6086.

[684] D. Ji, H. Wang, M. Tao, J. Huang, X.-S. Hua, and H. Lu, "Structural and statistical texture knowledge distillation for semantic segmentation," in *CVPR*. IEEE, 2022, pp. 16 855–16 864.

[685] D. Kang, P. Koniusz, M. Cho, and N. Murray, "Distilling self-supervised vision transformers for weakly-supervised few-shot classification 'i&' segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 19 627–19 638.

[686] S. Wu, W. Zhang, L. Xu, S. Jin, X. Li, W. Liu, and C. C. Loy, "Clipself: Vision transformer distills itself for open-vocabulary dense prediction," in *International Conference on Learning Representations (ICLR)*, 2024.

[687] M. Li, M. Halstead, and C. Mccool, "Knowledge distillation for efficient instance semantic segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5432–5439.

[688] F. Shen, A. Gurram, Z. Liu, H. Wang, and A. Knoll, "Diga: Distil to generalize and then adapt for domain adaptive semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 15 866–15 877.

[689] T. Berrada, C. Couprie, K. Alahari, and J. Verbeek, "Guided distillation for semi-supervised instance segmentation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2024, pp. 464–472.

[690] H. Gao, J. Guo, G. Wang, and Q. Zhang, "Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 9903–9913.

[691] M.-H. Phan, T.-A. Ta, S. L. Phung, L. Tran-Thanh, and A. Bouzerdoum, "Class similarity weighted knowledge distillation for continual semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 16 845–16 854.

[692] J.-W. Xiao, C.-B. Zhang, J. Feng, X. Liu, J. van de Weijer, and M.-M. Cheng, "Endpoints weight fusion for class incremental semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 7204–7213.

[693] X. Zheng, Y. Luo, P. Zhou, and L. Wang, "Distilling efficient vision transformers from cnns for semantic segmentation," *Pattern Recognition*, vol. 158, p. 111029, 2025.

[694] J. Zhu, Y. Luo, X. Zheng, H. Wang, and L. Wang, "A good student is cooperative and reliable: Cnn-transformer collaborative learning for semantic segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 11 686–11 696.

[695] Z. Wang, X. Yu, X. Han, W. Yu, Z. Huang, J. Jiao, and Z. Han, "P2seg: Pointly-supervised segmentation via mutual distillation," in *International Conference on Learning Representations (ICLR)*, 2024.

[696] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai, "Efficient medical image segmentation based on knowledge distillation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3820–3831, 2021.

[697] J. Li, G. Chen, H. Mao, D. Deng, D. Li, J. Hao, Q. Dou, and P.-A. Heng, "Flat-aware cross-stage distilled framework for imbalanced medical image classification," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 13433. Springer, 2022, pp. 217–226.

[698] S. Ghosh, K. Yu, and K. Batmanghelich, "Distilling blackbox to interpretable models for efficient transfer learning," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 14221. Springer, 2023, pp. 628–638.

[699] H. Wang, C. Ma, J. Zhang, Y. Zhang, J. Avery, L. Hull, and G. Carneiro, "Learnable cross-modal knowledge distillation for multi-modal learning with missing modality," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 14223. Springer, 2023, pp. 216–226.

[700] W. Cao, J. Zhang, Y. Xia, T. C. W. Mok, Z. Li, Y. Ye, L. Lu, J. Zheng, Y. Tang, and L. Zhang, "Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 11 238–11 247.

[701] S. A. Nasser, N. Gupte, and A. Sethi, "Reverse knowledge distillation: Training a large model using a small one for retinal image matching on limited data," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2024, pp. 7763–7772.

[702] K. Chen, T. Qin, V. H. F. Lee, H. Yan, and H. Li, "Learning robust shape regularization for generalizable medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 7, pp. 2693–2706, 2024.

[703] Y. Zhou, S. Du, H. Li, J. Yao, Y. Zhang, and Y. Wang, "Reprogramming distillation for medical foundation models," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 15011. Springer, 2024, pp. 533–543.

[704] W. Dong, B. Du, and Y. Xu, "Shape-intensity knowledge distillation for robust medical image segmentation," *Frontiers of Computer Science*, vol. 19, no. 9, p. 199705, 2025.

[705] Y. Ye, J. Zhang, Z. Chen, and Y. Xia, "Desd: Self-supervised learning with deep self-distillation for 3d medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 13434. Springer, 2022, pp. 545–555.

[706] J. Yi, Q. Bi, H. Zheng, H. Zhan, W. Ji, Y. Huang, S. Li, Y. Li, Y. Zheng, and F. Huang, "Hallucinated style distillation for single domain generalization in medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 15010. Springer, 2024, pp. 438–448.

[707] Y. Feng, B. Zhang, L. Xiao, Y. Yang, T. Gegen, and Z. Chen, "Enhancing medical imaging with gans synthesizing realistic images from limited data," in *IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)*, 2024, pp. 1192–1197.

[708] Y. Wang, P. Cao, Q. Hou, L. Lan, J. Yang, X. Liu, and O. R. Zaïane, "Progressively correcting soft labels via teacher team for knowledge distillation in medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 15009. Springer, 2024, pp. 521–530.

[709] Y. Xie, Y. Yin, Q. Li, and Y. Wang, "Deep mutual distillation for semi-supervised medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 14222. Springer, 2023, pp. 540–550.

[710] L. Zhong, X. Liao, S. Zhang, and G. Wang, "Semi-supervised pathological image segmentation via cross distillation of multiple attentions," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 14225. Springer, 2023, pp. 570–579.

[711] L. Huang, Y. Liang, and J. Liu, "Des-sam: Distillation-enhanced semantic sam for cervical nuclear segmentation with box annotation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 15009. Springer, 2024, pp. 223–234.

[712] C. You, Y. Zhou, R. Zhao, L. Staib, and J. S. Duncan, "Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 9, pp. 2228–2237, 2022.

[713] S. Sharma, A. Kumar, and J. Chandra, "Confidence matters: Enhancing medical image classification through uncertainty-driven contrastive self-distillation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 15010. Springer, 2024, pp. 133–142.

[714] O. S. El-Assiouti, G. Hamed, D. R. Khattab, and H. M. Ebeid, "Hdkd: Hybrid data-efficient knowledge distillation network for medical image classification," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109430, 2024.

[715] X. Liu, B. Hu, W. Huang, Y. Zhang, and Z. Xiong, "Efficient biomedical instance segmentation via knowledge distillation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 13434. Springer, 2022, pp. 14–24.

[716] X. Qi, Z. Wu, W. Zou, M. Ren, Y. Gao, M. Sun, S. Zhang, C. Shan, and Z. Sun, "Exploring generalizable distillation for efficient medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 7, pp. 4170–4183, 2024.

[717] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou, "General instance distillation for object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 2021, pp. 7842–7851.

[718] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, and C. Xu, "Distilling object detectors via decoupled features," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 2021, pp. 2154–2164.

[719] S. Guo, J. M. Álvarez, and M. Salzmann, "Distilling image classifiers in object detectors," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 1036–1047.

[720] Z. Kang, P. Zhang, X. Zhang, J. Sun, and N. Zheng, "Instance-conditional knowledge distillation for object detection," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 16 468–16 480.

[721] C. Yang, M. Ochal, A. J. Storkey, and E. J. Crowley, "Prediction-guided distillation for dense object detection," in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, vol. 13669. Springer, 2022, pp. 123–138.

[722] Y. Jang, W. Shin, J. Kim, S. S. Woo, and S.-H. Bae, "Glamd: Global and local attention mask distillation for object detectors," in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, vol. 13670. Springer, 2022, pp. 460–476.

[723] S. Lao, G. Song, B. Liu, Y. Liu, and Y. Yang, "Unikd: Universal knowledge distillation for mimicking homogeneous or heterogeneous

object detectors," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 6339–6349.

[724] Q. Lan and Q. Tian, "Gradient-guided knowledge distillation for object detectors," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2024, pp. 423–432.

[725] Z. Liu, X. Hu, and R. Nevatia, "Efficient feature distillation for zero-shot annotation object detection," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2024, pp. 882–891.

[726] D. Wu, P. Chen, X. Yu, G. Li, Z. Han, and J. Jiao, "Spatial self-distillation for object detection with inaccurate bounding boxes," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 6832–6842.

[727] Z. Jia, S. Sun, G. Liu, and B. Liu, "Mssd: Multi-scale self-distillation for object detection," *Visual Intelligence*, vol. 2, pp. 1–11, 2024.

[728] Z. Zheng, R. Ye, Q. Hou, D. Ren, P. Wang, W. Zuo, and M.-M. Cheng, "Localization distillation for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 070–10 083, 2023.

[729] L. Yang, X. Zhou, X. Li, L. Qiao, Z. Li, Z. Yang, G. Wang, and X. Li, "Bridging cross-task protocol inconsistency for distillation in dense object detection," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 17 129–17 138.

[730] Y. Jin, J. Wang, and D. Lin, "Multi-level logit distillation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 24 276–24 285.

[731] L. Li, J. Miao, D. Shi, W. Tan, Y. Ren, Y. Yang, and S. Pu, "Distilling detr with visual-linguistic knowledge for open-vocabulary object detection," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 6478–6487.

[732] H. Zhou, B. Qiao, L. Yang, J. Lai, and X. Xie, "Texture-guided saliency distilling for unsupervised salient object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 7257–7267.

[733] C. H. Nguyen, T. C. Nguyen, T. N. Tang, and N. L. H. Phan, "Improving object detection by label assignment distillation," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2022, pp. 1322–1331.

[734] Z. Ni, F. Yang, S. Wen, and G. Zhang, "Dual relation knowledge distillation for object detection," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023, pp. 1276–1284.

[735] R. Tang, Z. Liu, Y. Li, Y. Song, H. Liu, Q. Wang, J. Shao, G. Duan, and J. Tan, "Task-balanced distillation for object detection," *CoRR*, vol. abs/2208.03006, 2022.

[736] Z. Du, R. Zhang, M. Chang, X. Zhang, S. Liu, T. Chen, and Y. Chen, "Distilling object detectors with feature richness," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 5213–5224.

[737] L. Zhang and K. Ma, "Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors," in *International Conference on Learning Representations (ICLR)*, 2021.

[738] L. Yao, R. Pi, H. Xu, W. Zhang, Z. Li, and T. Zhang, "G-detkd: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 3571–3580.

[739] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, and C. Yuan, "Focal and global knowledge distillation for detectors," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 4633–4642.

[740] G. Li, X. Li, Y. Wang, S. Zhang, Y. Wu, and D. Liang, "Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation," in *AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, 2022, pp. 1306–1313.

[741] L. Zhang and K. Ma, "Structured knowledge distillation for accurate and efficient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 706–15 724, 2023.

[742] Y. Zhu, Q. Zhou, N. Liu, Z. Xu, Z. Ou, X. Mou, and J. Tang, "Scalekd: Distilling scale-aware knowledge in small object detector," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 19 723–19 733.

[743] P. Zhang, Z. Kang, T. Yang, X. Zhang, N. Zheng, and J. Sun, "Lgd: Label-guided self-distillation for object detection," in *AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, 2022, pp. 3309–3317.

[744] J. Wang, Y. Chen, Z. Zheng, X. Li, M.-M. Cheng, and Q. Hou, "Crosskd: Cross-head knowledge distillation for object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 19 723–19 733.

[745] X. Hou, J. Xing, Y. Qian, Y. Guo, S. Xin, J. Chen, K. Tang, M. Wang, Z. Jiang, L. Liu, and Y. Liu, "Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 26 541–26 551.

[746] W. Zhang, W. Deng, Z. Cui, J. Liu, and L. Jiao, "Object knowledge distillation for joint detection and tracking in satellite videos," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[747] F. R. Valverde, J. V. Hurtado, and A. Valada, "There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 607–11 616, 2021.

[748] X. Wang, S. Wang, C. Tang, L. Zhu, B. Jiang, Y. Tian, and J. Tang, "Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 19 248–19 257.

[749] X. Nie, Y. Li, L. Luo, N. Zhang, and J. Feng, "Dynamic kernel distillation for efficient pose estimation in videos," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 6941–6949.

[750] S. Chen, Y. Zhang, S. Huang, R. Yi, K. Fan, R. Zhang, P. Chen, J. Wang, S. Ding, and L. Ma, "Sdpose: Tokenized pose estimation via circulation-guide self-distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1082–1090.

[751] S. Guo, Y. Hu, J. M. Álvarez, and M. Salzmann, "Knowledge distillation for 6d pose estimation by aligning distributions of local predictions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 18 633–18 642.

[752] S. Ye, Y. Zhang, J. Hu, L. Cao, S. Zhang, L. Shen, J. Wang, S. Ding, and R. Ji, "Distilpose: Tokenized pose regression with heatmap distillation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 2163–2172.

[753] J. Wang, S. Jin, W. Liu, W. Liu, C. Qian, and P. Luo, "When human pose estimation meets robustness: Adversarial algorithms and benchmarks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 2021, pp. 11 855–11 864.

[754] N. Aghli and E. Ribeiro, "Combining weight pruning and knowledge distillation for cnn compression," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Computer Vision Foundation / IEEE, 2021, pp. 3191–3198.

[755] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2022.

[756] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 3517–3526.

[757] C. Li and G. H. Lee, "From synthetic to real: Unsupervised domain adaptation for animal pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 1482–1491.

[758] H. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7157–7173, 2023.

[759] Q. GUAN, Z. SHENG, and S. XUE, "Hrpose: Real-time high-resolution 6d pose estimation network using knowledge distillation," *Chinese Journal of Electronics*, vol. 32, no. 1, pp. 189–198, 2023.

[760] Z. Yang, A. Zeng, C. Yuan, and Y. Li, "Effective whole-body pose estimation with two-stages distillation," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 2023, pp. 4212–4222.

[761] Y. Jiang, C. Feng, F. Zhang, and D. Bull, "Mtkd: Multi-teacher knowledge distillation for image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, vol. 15097. Springer, 2024, pp. 364–382.

[762] H. Wang, Z. Wei, Q. Tang, S. Cheng, L. Wang, and Y. Li, "Attention guidance distillation network for efficient image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2024, pp. 6287–6296.

[763] Y. Wang and T. Zhang, "Osffnet: Omni-stage feature fusion network for lightweight image super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, 2024, pp. 5660–5668.

[764] M. Noroozi, I. Hadji, B. Martínez, A. Bulat, and G. Tzimiropoulos, "You only need one step: Fast super-resolution with stable diffusion via scale distillation," in *ECCV (29)*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., vol. 15087. Springer, 2024, pp. 145–161.

[765] H. Park, "Semantic super-resolution via self-distillation and adversarial learning," *IEEE Access*, vol. 12, pp. 2361–2370, 2024.

[766] S. Angarano, F. Salvetti, M. Martini, and M. Chiaberge, "Generative adversarial super-resolution at the edge with knowledge distillation," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106407, 2022.

[767] Y. Zhang, S. Lee, and A. Yao, "Pairwise distance distillation for unsupervised real-world image super-resolution," in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, vol. 15087. Springer, 2024, pp. 429–446.

[768] J. Xie, L. Gong, S. Shao, S. Lin, and L. Luo, "Hybrid knowledge distillation from intermediate layers for efficient single image super-resolution," *Neurocomputing*, vol. 554, p. 126592, 2023.

[769] H. Fang, Y. Long, X. Hu, Y. Ou, Y. Huang, and H. Hu, "Dual cross knowledge distillation for image super-resolution," *Journal of Visual Communication and Image Representation*, vol. 95, p. 103858, 2023.

[770] Y. Wang, S. Lin, Y. Qu, H. Wu, Z. Zhang, Y. Xie, and A. Yao, "Towards compact single image super-resolution via contrastive self-distillation," in *IJCAI*, Z.-H. Zhou, Ed. ijcai.org, 2021, pp. 1122–1128.

[771] Y. Zhang, H. Chen, X. Chen, Y. Deng, C. Xu, and Y. Wang, "Data-free knowledge distillation for image super-resolution," in *CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 7852–7861.

[772] L. Yu, X. Li, Y. Li, T. Jiang, Q. Wu, H. Fan, and S. Liu, "Dipnet: Efficiency distillation and iterative pruning for image super-resolution," in *CVPR Workshops*. IEEE, 2023, pp. 1692–1701.

[773] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A. C. Kot, and B. Wen, "Sinsr: Diffusion-based image super-resolution in a single step," in *CVPR*. IEEE, 2024, pp. 25 796–25 805.

[774] P. Zhao, L. Xie, J. Wang, Y. Zhang, and Q. Tian, "Progressive privileged knowledge distillation for online action detection," *Pattern Recognition*, vol. 129, p. 108741, 2022.

[775] X. Huang, H. Zhou, K. Yao, and K. Han, "Froster: Frozen clip is a strong teacher for open-vocabulary action recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[776] J. Hong, M. Fisher, M. Gharbi, and K. Fatahalian, "Video pose distillation for few-shot, fine-grained sports action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2021, pp. 9234–9243.

[777] J. Guo, H. Liu, S. Sun, T. Guo, M. Zhang, and C. Si, "Fsar: Federated skeleton-based action recognition with adaptive topology structure and knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2023, pp. 10 366–10 376.

[778] G. Radevski, D. Grujicic, M. B. Blaschko, M.-F. Moens, and T. Tuytelaars, "Multimodal distillation for egocentric action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 5190–5201.

[779] R. Dai, S. Das, and F. Brémond, "Learning an augmented rgb representation with cross-modal knowledge distillation for action detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2021, pp. 13 033–13 044.

[780] C. Bian, W. Feng, L. Wan, and S. Wang, "Structural knowledge distillation for efficient skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2963–2976, 2021.

[781] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2020, pp. 11 165–11 172.

[782] A. Porrello, L. Bergamini, and S. Calderara, "Robust re-identification by multiple views knowledge distillation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 12355. Springer, 2020, pp. 93–110.

[783] M. Budnik and Y. Avrithis, "Asymmetric metric learning for knowledge transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 2021, pp. 8228–8238.

[784] H. Wu, M. Wang, W. Zhou, H. Li, and Q. Tian, "Contextual similarity distillation for asymmetric image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 9479–9488.

[785] K. Xu, X. Zou, and J. Zhou, "Lstkc: Long short-term knowledge consolidation for lifelong person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2024, pp. 16 202–16 210.

[786] J. Tian, X. Xu, Z. Wang, F. Shen, and X. Liu, "Relationship-preserving knowledge distillation for zero-shot sketch based image retrieval," in *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*. ACM, 2021, pp. 5473–5481.

[787] Y. Xie, H. Wu, Y. Lin, J. Zhu, and H. Zeng, "Pairwise difference relational distillation for object re-identification," *Pattern Recognition*, vol. 152, p. 110455, 2024.

[788] Y. Xie, H. Zhang, X. Xu, J. Zhu, and S. He, "Towards a smaller student: Capacity dynamic distillation for efficient image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023, pp. 16 006–16 015.

[789] P. Suma and G. Tolias, "Large-to-small image resolution asymmetry in deep metric learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2023, pp. 1451–1460.

[790] Y. Xie, Y. Lin, W. Cai, X. Xu, H. Zhang, Y. Du, and S. He, "D3still: Decoupled differential distillation for asymmetric image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 17 181–17 190.

[791] Y.-C. Hsu, J. Smith, Y. Shen, Z. Kira, and H. Jin, "A closer look at knowledge distillation with features, logits, and gradients," *arXiv preprint arXiv:2203.10163*, 2022.

[792] U. Ojha, Y. Li, A. Sundara Rajan, Y. Liang, and Y. J. Lee, "What knowledge gets distilled in knowledge distillation?" *Advances in Neural Information Processing Systems*, vol. 36, pp. 11 037–11 048, 2023.

[793] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[794] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[795] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," in *ICLR*, 2021.

[796] T. Han, S. Huang, Z. Ding, W. Sun, Y. Feng, C. Fang, J. Li, H. Qian, C. Wu, Q. Zhang *et al.*, "On the effectiveness of distillation in mitigating backdoors in pre-trained encoder," *arXiv preprint arXiv:2403.03846*, 2024.

[797] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 902–14 912.

[798] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros, "Dataset distillation," *arXiv preprint arXiv:1811.10959*, 2018.