Knowledge Distillation: A Survey

Jianping Gou[1,2] · Baosheng Yu[2] · Stephen J. Maybank[3] · Dacheng Tao[2]

Abstract

In recent years, deep neural networks have been successful in both industry and academia, especially for computer vision tasks. The great success of deep learning is mainly due to its scalability to encode large-scale data and to maneuver billions of model parameters. However, it is a challenge to deploy these cumbersome deep models on devices with limited resources, e.g., mobile phones and embedded devices, not only because of the high computational complexity but also the large storage requirements. To this end, a variety of model compression and acceleration techniques have been developed. As a representative type of model compression and acceleration, knowledge distillation effectively learns a small student model from a large teacher model. It has received rapid increasing attention from the community. This paper provides a comprehensive survey of knowledge distillation from the perspectives of knowledge categories, training schemes, teacher–student architecture, distillation algorithms, performance comparison and applications. Furthermore, challenges in knowledge distillation are briefly reviewed and comments on future research are discussed and forwarded.

1 Introduction

During the last few years, deep learning has been the basis of many successes in artificial intelligence, including a variety of applications in computer vision (Krizhevsky et al. 2012), reinforcement learning (Silver et al. 2016; Ashok et al. 2018; Lai et al. 2020), and natural language processing (Devlin et al.

✉ Dacheng Tao
  dacheng.tao@sydney.edu.au

  Jianping Gou
  cherish.gjp@gmail.com

  Baosheng Yu
  baosheng.yu@sydney.edu.au

  Stephen J. Maybank
  sjmaybank@dcs.bbk.ac.uk

[1] School of Computer Science and Communication Engineering and Jiangsu Key Laboratory of Security Technology for Industrial Cyberspace, Jiangsu University, Zhenjiang 212013, China

[2] School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia

[3] Department of Computer Science and Information Systems, Birkbeck College, University of London, London, UK

2019). With the help of many recent techniques, including residual connections (He et al. 2016, 2020b) and batch normalization (Ioffe and Szegedy 2015), it is easy to train very deep models with thousands of layers on powerful GPU or TPU clusters. For example, it takes less than ten minutes to train a ResNet model on a popular image recognition benchmark with millions of images (Deng et al. 2009; Sun et al. 2019); It takes no more than one and a half hours to train a powerful BERT model for language understanding (Devlin et al. 2019; You et al. 2019). The large-scale deep models have achieved overwhelming successes, however the huge computational complexity and massive storage requirements make it a great challenge to deploy them in real-time applications, especially on devices with limited resources, such as video surveillance and autonomous driving cars.

To develop efficient deep models, recent works usually focus on 1) efficient building blocks for deep models, including depthwise separable convolution, as in MobileNets (Howard et al. 2017; Sandler et al. 2018) and ShuffleNets (Zhang et al. 2018a; Ma et al. 2018); and 2) model compression and acceleration techniques, in the following categories (Cheng et al. 2018).

- Parameter pruning and sharing: These methods focus on removing inessential parameters from deep neural net-
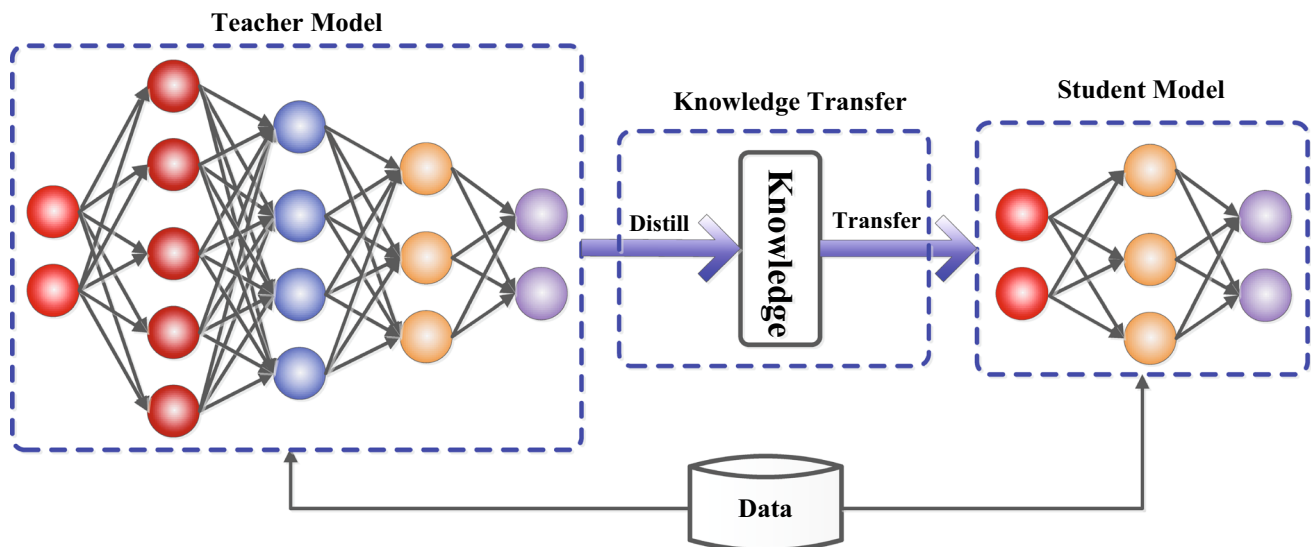
Fig. 1 The generic teacher–student framework for knowledge distillation

works without any significant effect on the performance. This category is further divided into model quantization (Wu et al. 2016), model binarization (Courbariaux et al. 2015), structural matrices (Sindhwani et al. 2015) and parameter sharing (Han et al. 2015; Wang et al. 2019f).

- Low-rank factorization: These methods identify redundant parameters of deep neural networks by employing the matrix and tensor decomposition (Yu et al. 2017; Denton et al. 2014).
- Transferred compact convolutional filters: These methods remove inessential parameters by transferring or compressing the convolutional filters (Zhai et al. 2016).
- Knowledge distillation (KD): These methods distill the knowledge from a larger deep neural network into a small network (Hinton et al. 2015).

A comprehensive review on model compression and acceleration is outside the scope of this paper. The focus of this paper is knowledge distillation, which has received increasing attention from the research community in recent years. Large deep neural networks have achieved remarkable success with good performance, especially in the real-world scenarios with large-scale data, because the over parameterization improves the generalization performance when new data is considered (Zhang et al. 2018; Brutzkus and Globerson 2019; Allen-Zhu et al. 2019; Arora et al. 2018; Tu et al. 2020). However, the deployment of deep models in mobile devices and embedded systems is a great challenge, due to the limited computational capacity and memory of the devices. To address this issue, Bucilua et al. (2006) first proposed model compression to transfer the information from a large model or an ensemble of models into training a small model without a significant drop in accuracy. The learning of a small model from a large model is later formally popularized as knowledge distillation (Hinton et al. 2015). In knowledge distillation, a small student model is generally supervised by a large teacher model (Bucilua et al. 2006; Ba and Caruana 2014; Hinton et al. 2015; Urban et al. 2017). The main idea is that the student model mimics the teacher model in order to obtain a competitive or even a superior performance. The key problem is how to transfer the knowledge from a large teacher model to a small student model. Basically, a knowledge distillation system is composed of three key components: knowledge, distillation algorithm, and teacher–student architecture. A general teacher–student framework for knowledge distillation is shown in Fig. 1.

Although the great success in practice, there are not too many works on either the theoretical or empirical understanding of knowledge distillation (Cheng et al. 2020; Phuong and Lampert 2019a; Cho and Hariharan 2019). Specifically, to understand the working mechanisms of knowledge distillation, Phuong & Lampert obtained a theoretical justification for a generalization bound with fast convergence of learning distilled student networks in the scenario of deep linear classifiers (Phuong and Lampert 2019a). This justification answers what and how fast the student learns and reveals the factors of determining the success of distillation. Successful distillation relies on data geometry, optimization bias of distillation objective and strong monotonicity of the student classifier. Cheng et al. quantified the extraction of visual concepts from the intermediate layers of a deep neural network, to explain knowledge distillation (Cheng et al. 2020). Ji and Zhu (2020) theoretically explained knowledge distillation on a wide neural network from the respective of risk bound, data efficiency and imperfect teacher. Cho & Hariharan empirically analyzed in detail the efficacy of knowledge distillation
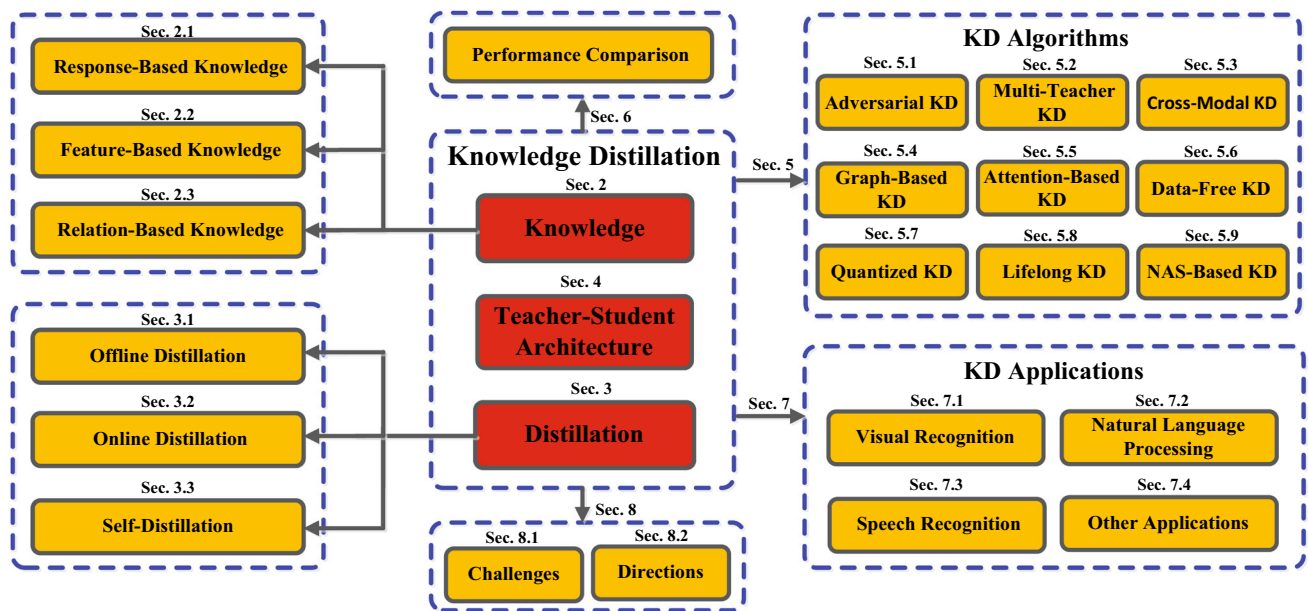
Fig. 2 The schematic structure of knowledge distillation and the relationship between the adjacent sections. The body of this survey mainly contains the fundamentals of knowledge distillation, knowledge types, distillation schemes, teacher–student architecture, distillation algorithms, performance comparison, applications, discussions, challenges, and future directions. Note that 'Section' is abbreviated as 'Sec.' in this figure

(Cho and Hariharan 2019). Empirical results show that a larger model may not be a better teacher because of model capacity gap (Mirzadeh et al. 2020). Experiments also show that distillation adversely affects the student learning. The empirical evaluation of different forms of knowledge distillation about knowledge, distillation and mutual affection between teacher and student is not covered by Cho and Hariharan (2019). Knowledge distillation has also been explored for label smoothing, for assessing the accuracy of the teacher and for obtaining a prior for the optimal output layer geometry (Tang et al. 2020).

Knowledge distillation for model compression is similar to the way in which human beings learn. Inspired by this, recent knowledge distillation methods have extended to teacher–student learning (Hinton et al. 2015), mutual learning (Zhang et al. 2018b), assistant teaching (Mirzadeh et al. 2020), lifelong learning (Zhai et al. 2019), and self-learning (Yuan et al. 2020). Most of the extensions of knowledge distillation concentrate on compressing deep neural networks. The resulting lightweight student networks can be easily deployed in applications such as visual recognition, speech recognition, and natural language processing (NLP). Furthermore, the knowledge transfer from one model to another in knowledge distillation can be extended to other tasks, such as adversarial attacks (Papernot et al. 2016), data augmentation (Lee et al. 2019a; Gordon and Duh 2019), data privacy and security (Wang et al. 2019a). Motivated by knowledge distillation for model compression, the idea of knowledge

transfer has been further applied in compressing the training data, i.e., dataset distillation, which transfers the knowledge from a large dataset into a small dataset to reduce the training loads of deep models (Wang et al. 2018c; Bohdal et al. 2020).

In this paper, we present a comprehensive survey on knowledge distillation. The main objectives of this survey are to 1) provide an overview on knowledge distillation, including several typical knowledge, distillation and architectures; 2) review the recent progress of knowledge distillation, including algorithms and applications to different real-world scenarios; and 3) address some hurdles and provide insights to knowledge distillation based on different perspectives of knowledge transfer, including different types of knowledge, training schemes, distillation algorithms and structures, and applications. Recently, there is also a similar survey on knowledge distillation (Wang and Yoon. 2020), which presents the comprehensive progress from different perspective of teacher–student learning for vision and its challenges. Different from Wang and Yoon. (2020), our survey mainly focuses on knowledge distillation from a wide perspective of knowledge types, distillation schemes, distillation algorithms, performance comparison and different application areas.

The organization of this paper is shown in Fig. 2. The different kinds of knowledge and of distillation are summarized in Sects. 2 and 3, respectively. The existing studies about the teacher–student structures in knowledge distillation are illustrated in Sect. 4. The latest knowledge distillation approaches

Fig. 3 The schematic illustrations of sources of response-based knowledge, feature-based knowledge and relation-based knowledge in a deep teacher network
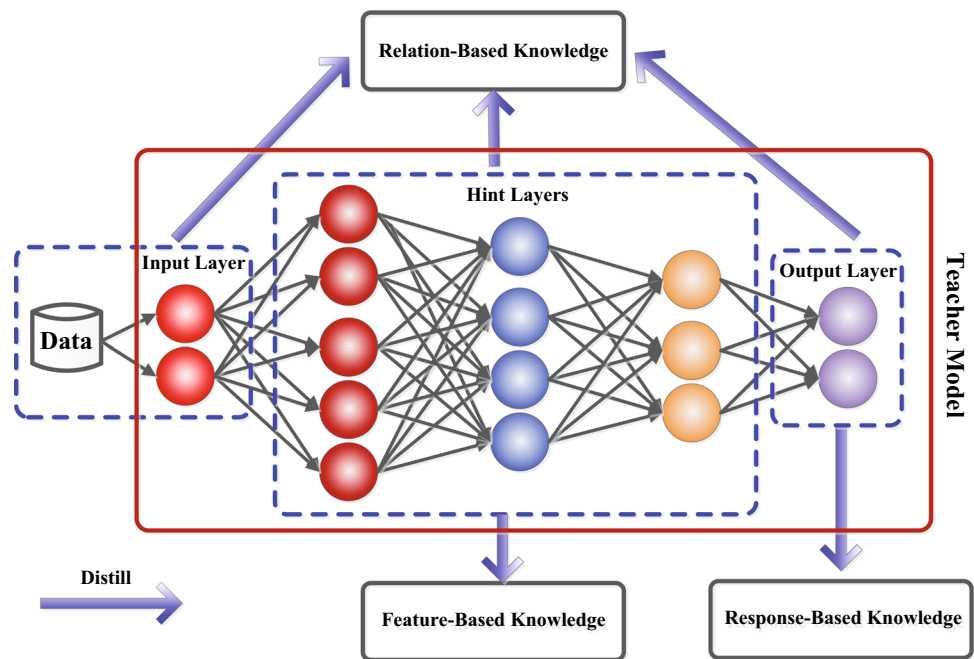
are comprehensively summarized in Sect. 5. The performance comparison of knowledge distillation is reported in Sect. 6. The many applications of knowledge distillation are illustrated in Sect. 7. Challenging problems and future directions in knowledge distillation are discussed and conclusion is given in Sect. 8.

2 Knowledge

In knowledge distillation, knowledge types, distillation strategies and the teacher–student architectures play the crucial role in the student learning. In this section, we focus on different categories of knowledge for knowledge distillation. A vanilla knowledge distillation uses the logits of a large deep model as the teacher knowledge (Hinton et al. 2015; Kim et al. 2018; Ba and Caruana 2014; Mirzadeh et al. 2020). The activations, neurons or features of intermediate layers also can be used as the knowledge to guide the learning of the student model (Romero et al. 2015; Huang and Wang 2017; Ahn et al. 2019; Heo et al. 2019c; Zagoruyko and Komodakis 2017). The relationships between different activations, neurons or pairs of samples contain rich information learned by the teacher model (Yim et al. 2017; Lee and Song 2019; Liu et al. 2019g; Tung and Mori 2019; Yu et al. 2019). Furthermore, the parameters of the teacher model (or the connections between layers) also contain another knowledge (Liu et al. 2019c). We discuss different forms of knowledge in the following categories: response-based knowledge, feature-based knowledge, and relation-based knowledge. An
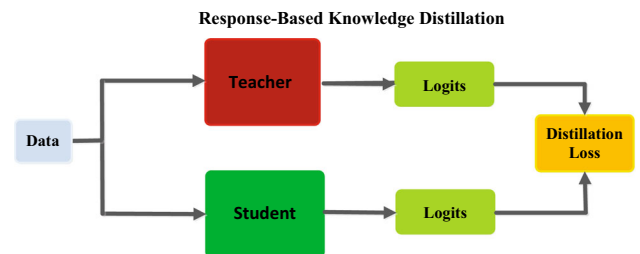

Fig. 4 The generic response-based knowledge distillation

intuitive example of different categories of knowledge within a teacher model is shown in Fig. 3.

2.1 Response-Based Knowledge

Response-based knowledge usually refers to the neural response of the last output layer of the teacher model. The main idea is to directly mimic the final prediction of the teacher model. The response-based knowledge distillation is simple yet effective for model compression, and has been widely used in different tasks and applications. Given a vector of **logits** z as the outputs of the last fully connected layer of a deep model, the distillation loss for response-based knowledge can be formulated as

$$L_{ResD}(z_t, z_s) = \mathcal{L}_R(z_t, z_s), \tag{1}$$

where $\mathcal{L}_R(.)$ indicates the divergence loss of logits, and z_t and z_s are logits of teacher and student, respectively. A typical response-based KD model is shown in Fig. 4. The response-based knowledge can be used for different types
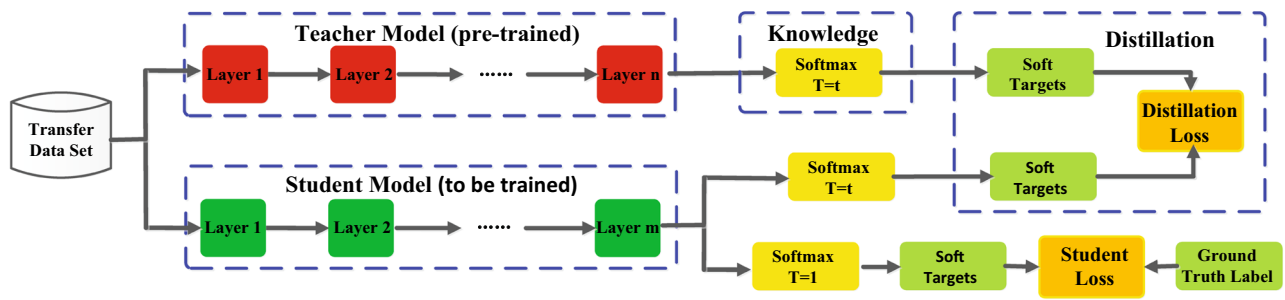
Fig. 5 The specific architecture of the benchmark knowledge distillation (Hinton et al. 2015)

of model predictions. For example, the response in object detection task may contain the logits together with the offset of a bounding box (Chen et al. 2017). In semantic landmark localization tasks, e.g., human pose estimation, the response of the teacher model may include a heatmap for each landmark (Zhang et al. 2019a). Recently, response-based knowledge has been further explored to address the information of ground-truth label as the conditional targets (Meng et al. 2019).

The most popular response-based knowledge for image classification is known as soft targets (Hinton et al. 2015; Ba and Caruana 2014). Specifically, soft targets are the probabilities that the input belongs to the classes and can be estimated by a softmax function as

$$p(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}, \tag{2}$$

where z_i is the logit for the i-th class, and a **temperature** factor T is introduced to control the importance of each soft target. As stated in Hinton et al. (2015), soft targets contain the informative dark knowledge from the teacher model. Accordingly, the distillation loss for soft logits can be rewritten as

$$L_{ResD}(p(z_t, T), p(z_s, T)) = \mathcal{L}_R(p(z_t, T), p(z_s, T)). \tag{3}$$

Generally, $\mathcal{L}_R(p(z_t, T), p(z_s, T))$ often employs Kullback-Leibler divergence loss. Clearly, optimizing Eq. (1) or (3) can make the logits z_s of student match the ones z_t of teacher. To easily understand the response-based knowledge distillation, the benchmark model of a vanilla knowledge distillation, which is the joint of the distillation and student losses, is given in Fig. 5. Note that the student loss is always defined as the cross-entropy loss $\mathcal{L}_{CE}(y, p(z_s, T = 1))$ between the ground truth label and the soft logits of the student model.

The idea of the response-based knowledge is straightforward and easy to understand, especially in the context of "dark knowledge". From another perspective, the effectiveness of the soft targets is analogous to label smoothing (Kim and Kim 2017) or regularizers (Muller et al. 2019; Ding
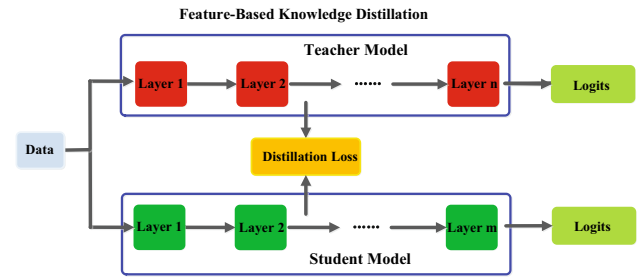

Fig. 6 The generic feature-based knowledge distillation

et al. 2019). However, the response-based knowledge usually relies on the output of the last layer, e.g., soft targets, and thus fails to address the intermediate-level supervision from the teacher model, which turns out to be very important for representation learning using very deep neural networks (Romero et al. 2015). Since the soft logits are in fact the class probability distribution, the response-based knowledge distillation is also limited to the supervised learning.

2.2 Feature-Based Knowledge

Deep neural networks are good at learning multiple levels of feature representation with increasing abstraction. This is known as representation learning (Bengio et al. 2013). Therefore, both the output of the last layer and the output of intermediate layers, i.e., feature maps, can be used as the knowledge to supervise the training of the student model. Specifically, feature-based knowledge from the intermediate layers is a good extension of response-based knowledge, especially for the training of thinner and deeper networks.

The intermediate representations were first introduced in Fitnets (Romero et al. 2015), to provide hints [1] to improve the training of the student model. The main idea is to directly match the feature activations of the teacher and the student. Inspired by this, a variety of other methods have been proposed to match the features indirectly (Zagoruyko and Komodakis 2017; Kim et al. 2018; Heo et al. 2019c; Passban

---

[1] A hint means the output of a teacher's hidden layer that supervises the student's learning.

Table 1 A summary of feature-based knowledge

Methods	Knowledge types	Knowledge sources	Distillation losses
Fitnet (Romero et al. 2015)	Feature representation	Hint layer	$\mathcal{L}_2(.)$
NST (Huang and Wang 2017)	Neuron selectivity patterns	Hint layer	$\mathcal{L}_{MMD}(.)$
AT (Zagoruyko and Komodakis 2017)	Attention maps	Multi-layer group	$\mathcal{L}_2(.)$
FT (Kim et al. 2018)	Paraphraser	Multi-layer group	$\mathcal{L}_1(.)$
Rocket Launching (Zhou et al. 2018)	Sharing parameters	Hint layer	$\mathcal{L}_2(.)$
KR (Liu et al. 2019c)	Parameters distribution	Multi-layer group	$\mathcal{L}_{CE}(.)$
AB (Heo et al. 2019c)	Activation boundaries	Pre-ReLU	$\mathcal{L}_2(.)$
Shen et al. (2019a)	Knowledge amalgamation	Hint layer	$\mathcal{L}_2(.)$
Heo et al. (2019a)	Margin ReLU	Pre-ReLU	$\mathcal{L}_2(.)$
FN (Xu et al. 2020b)	Feature representation	Fully-connected layer	$\mathcal{L}_{CE}(.)$
DFA (Guan et al. 2020)	Feature aggregation	Hint layer	$\mathcal{L}_2(.)$
AdaIN (Yang et al. 2020a)	Feature statistics	Hint layer	$\mathcal{L}_2(.)$
FN (Xu et al. 2020b)	Feature representation	Penultimate layer	$\mathcal{L}_{CE}(.)$
EC-KD (Wang et al. 2020b)	Feature representation	Hint layer	$\mathcal{L}_2(.)$
ALP-KD (Passban et al. 2021)	Attention-based layer projection	Hint layer	$\mathcal{L}_2(.)$
SemCKD (Chen et al. 2021)	Feature maps	Hint layer	$\mathcal{L}_2(.)$

et al. 2021; Chen et al. 2021; Wang et al. 2020b). To be specific, Zagoruyko and Komodakis (2017) derived an "attention map" from the original feature maps to express knowledge. The attention map was generalized by Huang and Wang (2017) using neuron selectivity transfer. Passalis and Tefas (2018) transferred knowledge by matching the probability distribution in feature space. To make it easier to transfer the teacher knowledge, Kim et al. (2018) introduced so called "factors" as a more understandable form of intermediate representations. To reduce the performance gap between teacher and student, Jin et al. (2019) proposed route constrained hint learning, which supervises student by outputs of hint layers of teacher. Recently, Heo et al. (2019c) proposed to use the activation boundary of the hidden neurons for knowledge transfer. Interestingly, the parameter sharing of intermediate layers of the teacher model together with response-based knowledge is also used as the teacher knowledge (Zhou et al. 2018). To match the semantics between teacher and student, Chen et al. (2021) proposed cross-layer knowledge distillation, which adaptively assigns proper teacher layers for each student layer via attention allocation.

Generally, the distillation loss for feature-based knowledge transfer can be formulated as

$$L_{FeaD}\big(f_t(x), f_s(x)\big) = \mathcal{L}_F\big(\Phi_t(f_t(x)), \Phi_s(f_s(x))\big), \quad (4)$$

where $f_t(x)$ and $f_s(x)$ are the feature maps of the intermediate layers of teacher and student models, respectively. The transformation functions, $\Phi_t(f_t(x))$ and $\Phi_s(f_s(x))$, are usually applied when the feature maps of teacher and student models are not in the same shape. $\mathcal{L}_F(.)$ indicates the similarity function used to match the feature maps of teacher and student models. A general feature-based KD model is shown in Fig. 6. We also summarize different types of feature-based knowledge in Table 1 from the perspective of feature types, source layers, and distillation loss. Specifically, $\mathcal{L}_2(.)$, $\mathcal{L}_1(.)$, $\mathcal{L}_{CE}(.)$ and $\mathcal{L}_{MMD}(.)$ indicate l_2-norm distance, l_1-norm distance, cross-entropy loss and maximum mean discrepancy loss, respectively. Though feature-based knowledge transfer provides favorable information for the learning of the student model, how to effectively choose the hint layers from the teacher model and the guided layers from the student model remains to be further investigated (Romero et al. 2015). Due to the significant differences between sizes of hint and guided layers, how to properly match feature representations of teacher and student also needs to be explored.

2.3 Relation-Based Knowledge

Both response-based and feature-based knowledge use the outputs of specific layers in the teacher model. Relation-based knowledge further explores the relationships between different layers or data samples.

To explore the relationships between different feature maps, Yim et al. (2017) proposed a flow of solution process (FSP), which is defined by the Gram matrix between two layers. The FSP matrix summarizes the relations between pairs of feature maps. It is calculated using the inner products between features from two layers. Using the correlations between feature maps as the distilled knowledge, knowledge distillation via singular value decomposition was proposed to extract key information in the feature maps (Lee et al.

2018). To use the knowledge from multiple teachers, Zhang and Peng (2018) formed two graph by respectively using the logits and features of each teacher model as the nodes. Specifically, the importance and relationships of the different teachers are modeled by the logits and representation graphs before the knowledge transfer (Zhang and Peng 2018). Multi-head graph-based knowledge distillation was proposed by Lee and Song (2019). The graph knowledge is the intra-data relations between any two feature maps via multi-head attention network. To explore the pairwise hint information, the student model also mimics the mutual information flow from pairs of hint layers of the teacher model (Passalis et al. 2020b). In general, the distillation loss of relation-based knowledge based on the relations of feature maps can be formulated as

$$L_{RelD}(f_t, f_s) = \mathcal{L}_{R^1}\left(\Psi_t\left(\hat{f}_t, \check{f}_t\right), \Psi_s\left(\hat{f}_s, \check{f}_s\right)\right), \qquad (5)$$

where f_t and f_s are the feature maps of teacher and student models, respectively. Pairs of feature maps are chosen from the teacher model, $\hat{f}_t$ and $\check{f}_t$, and from the student model, $\hat{f}_s$ and $\check{f}_s$. $\Psi_t(.)$ and $\Psi_s(.)$ are the similarity functions for pairs of feature maps from the teacher and student models. $\mathcal{L}_{R^1}(.)$ indicates the correlation function between the teacher and student feature maps.

Traditional knowledge transfer methods often involve individual knowledge distillation. The individual soft targets of a teacher are directly distilled into student. In fact, the distilled knowledge contains not only feature information but also mutual relations of data samples (You et al. 2017; Park et al. 2019). Specifically, Liu et al. (2019g) proposed a robust and effective knowledge distillation method via instance relationship graph. The transferred knowledge in instance relationship graph contains instance features, instance relationships and the feature space transformation cross layers. Park et al. (2019) proposed a relational knowledge distillation, which transfers the knowledge from instance relations. Based on idea of manifold learning, the student network is learned by feature embedding, which preserves the feature similarities of samples in the intermediate layers of the teacher networks (Chen et al. 2021). The relations between data samples are modelled as probabilistic distribution using feature representations of data (Passalis and Tefas 2018; Passalis et al. 2020a). The probabilistic distributions of teacher and student are matched by knowledge transfer. (Tung and Mori 2019) proposed a similarity-preserving knowledge distillation method. In particular, similarity-preserving knowledge, which arises from the similar activations of input pairs in the teacher networks, is transferred into the student network, with the pairwise similarities preserved. Peng et al. (2019a) proposed a knowledge distillation method based on correlation congruence, in
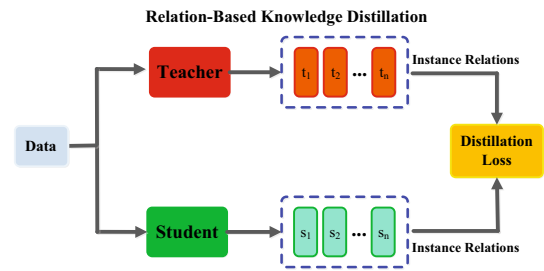

Fig. 7 The generic instance relation-based knowledge distillation

which the distilled knowledge contains both the instance-level information and the correlations between instances. Using the correlation congruence for distillation, the student network can learn the correlation between instances.

As described above, the distillation loss of relation-based knowledge based on the instance relations can be formulated as

$$L_{RelD}(F_t, F_s) = \mathcal{L}_{R^2}\big(\psi_t(t_i, t_j), \psi_s(s_i, s_j)\big), \qquad (6)$$

where $(t_i, t_j) \in F_t$ and $(s_i, s_j) \in F_s$, and F_t and F_s are the sets of feature representations from the teacher and student models, respectively. $\psi_t(.)$ and $\psi_s(.)$ are the similarity functions of (t_i, t_j) and (s_i, s_j). $\mathcal{L}_{R^2}(.)$ is the correlation function between the teacher and student feature representations. A typical instance relation-based KD model is shown in Fig. 7.

Distilled knowledge can be categorized from different perspectives, such as structured knowledge of the data (Liu et al. 2019g; Chen et al. 2021; Peng et al. 2019a; Tung and Mori 2019; Tian et al. 2020), privileged information about input features (Lopez-Paz et al. 2016; Vapnik and Izmailov 2015). A summary of differnet categories of relation-based knowledge is shown in Table 2. Specifically, $\mathcal{L}_{EM}(.)$, $\mathcal{L}_H(.)$, $\mathcal{L}_{AW}(.)$ and $\|.\|_F$ are Earth Mover distance, Huber loss, Angle-wise loss and Frobenius norm, respectively. Although some types of relation-based knowledge are provided recently, how to model the relation information from feature maps or data samples as knowledge still deserves further study.

3 Distillation Schemes

In this section, we discuss the distillation schemes (i.e. training schemes) for both teacher and student models. According to whether the teacher model is updated simultaneously with the student model or not, the learning schemes of knowledge distillation can be directly divided into three main categories: **offline distillation**, **online distillation** and **self-distillation**, as shown in Fig. 8.

Table 2 A summary of relation-based knowledge

Methods	Knowledge types	Knowledge sources	Distillation losses
FSP (Yim et al. 2017)	FSP matrix	End of multi-layer group	$\mathcal{L}_2(.)$
You et al. (2017)	Instance relation	Hint layers	$\mathcal{L}_2(.)$
Zhang and Peng (2018)	Logits graph, Representation graph	Softmax layers, Hint layers	$\mathcal{L}_{EM}(.), \mathcal{L}_{MMD}(.)$
DarkRank (Chen et al. 2018c)	Similarity DarkRank	Fully-connected layers	$\mathcal{L}_{KL}(.)$
MHGD (Lee and Song 2019)	Multi-head graph	Hint layers	$\mathcal{L}_{KL}(.)$
RKD (Park et al. 2019)	Instance relation	Fully-connected layers	$\mathcal{L}_H(.), \mathcal{L}_{AW}(.)$
IRG (Liu et al. 2019g)	Instance relationship graph	Hint layers	$\mathcal{L}_2(.)$
SP (Tung and Mori 2019)	Similarity matrix	Hint layers	$\|.\|_F$
CCKD (Peng et al. 2019a)	Instance relation	Hint layers	$\mathcal{L}_2(.)$
MLKD (Yu et al. 2019)	Instance relation	Hint layers	$\|.\|_F$
PKT(Passalis et al. 2020a)	Similarity probability distribution	Fully-connected layers	$\mathcal{L}_{KL}(.)$
Passalis et al. (2020b)	Mutual information flow	Hint layers	$\mathcal{L}_{KL}(.)$
LP (Chen et al. 2021)	Instance relation	Hint layers	$\mathcal{L}_2(.)$
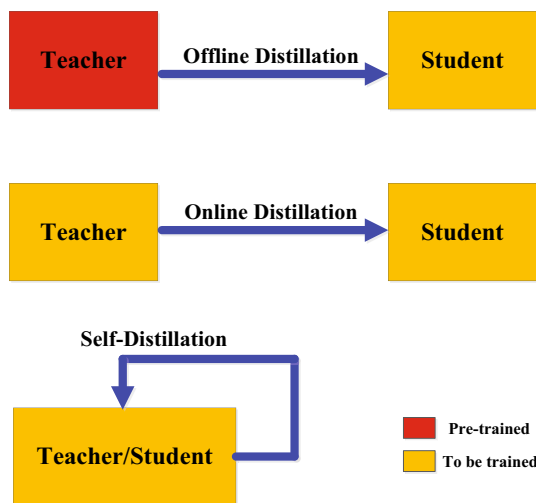
Fig. 8 Different distillations. The red color for "pre-trained" means networks are learned before distillation and the yellow color for "to be trained" means networks are learned during distillation

3.1 Offline Distillation

Most of previous knowledge distillation methods work offline. In vanilla knowledge distillation (Hinton et al. 2015), the knowledge is transferred from a pre-trained teacher model into a student model. Therefore, the whole training process has two stages, namely: 1) the large teacher model is first trained on a set of training samples before distillation; and 2) the teacher model is used to extract the knowledge in the forms of logits or the intermediate features, which are then used to guide the training of the student model during distillation.

The first stage in offline distillation is usually not discussed as part of knowledge distillation, i.e., it is assumed that the teacher model is pre-defined. Little attention is paid to the teacher model structure and its relationship with the student model. Therefore, the offline methods mainly focus on improving different parts of the knowledge transfer, including the design of knowledge (Hinton et al. 2015; Romero et al. 2015) and the loss functions for matching features or distributions matching (Huang and Wang 2017; Passalis and Tefas 2018; Zagoruyko and Komodakis 2017; Mirzadeh et al. 2020; Li et al. 2020d; Heo et al. 2019b; Asif et al. 2020). The main advantage of offline methods is that they are simple and easy to be implemented. For example, the teacher model may contain a set of models trained using different software packages, possibly located on different machines. The knowledge can be extracted and stored in a cache.

The offline distillation methods usually employ one-way knowledge transfer and two-phase training procedure. However, the complex high-capacity teacher model with huge training time can not be avoided, while the training of the student model in offline distillation is usually efficient under the guidance of the teacher model. Moreover, the capacity gap between large teacher and small student always exists, and student often largely relies on teacher.

3.2 Online Distillation

Although offline distillation methods are simple and effective, some issues in offline distillation have attracted increasing attention from the research community (Mirzadeh et al. 2020). To overcome the limitation of offline distillation, online distillation is proposed to further improve the performance of the student model, especially when a large-capacity high performance teacher model is not available (Zhang et al. 2018b; Chen et al. 2020a). In online distillation, both the teacher model and the student model are updated simulta-

neously, and the whole knowledge distillation framework is end-to-end trainable.

A variety of online knowledge distillation methods have been proposed, especially in the last few years (Zhang et al. 2018b; Chen et al. 2020a; Xie et al. 2019; Anil et al. 2018; Kim et al. 2019b; Zhou et al. 2018; Walawalkar et al. 2020; Wu and Gong 2021; Zhang et al. 2021a). Specifically, in deep mutual learning (Zhang et al. 2018b), multiple neural networks work in a collaborative way. Any one network can be the student model and other models can be the teacher during the training process. To improve generalization ability, deep mutual learning is extended by using ensemble of soft logits (Guo et al. 2020). Chen et al. (2020a) further introduced auxiliary peers and a group leader into deep mutual learning to form a diverse set of peer models. To reduce the computational cost, Zhu and Gong (2018) proposed a multi-branch architecture, in which each branch indicates a student model and different branches share the same backbone network. Rather than using the ensemble of logits, Kim et al. (2019b) introduced a feature fusion module to construct the teacher classifier. Xie et al. (2019) replaced the convolution layer with cheap convolution operations to form the student model. Anil et al. (2018) employed online distillation to train large-scale distributed neural network, and proposed a variant of online distillation called co-distillation. Co-distillation in parallel trains multiple models with the same architectures and any one model is trained by transferring the knowledge from the other models. Recently, an online adversarial knowledge distillation method is proposed to simultaneously train multiple networks by the discriminators using knowledge from both the class probabilities and a feature map (Chung et al. 2020). Adversarial co-distillation is lately devised by using GAN to generate divergent examples (Zhang et al. 2021a).

Online distillation is a one-phase end-to-end training scheme with efficient parallel computing. However, existing online methods (e.g., mutual learning) usually fails to address the high-capacity teacher in online settings, making it an interesting topic to further explore the relationships between the teacher and student model in online settings.

3.3 Self-Distillation

In self-distillation, the same networks are used for the teacher and the student models (Zhang et al. 2019b; Hou et al. 2019; Zhang and Sabuncu 2020; Yang et al. 2019b; Lee et al. 2019a; Phuong and Lampert 2019b; Lan et al. 2018; Xu and Liu 2019; Mobahi et al. 2020). This can be regarded as a special case of online distillation. Specifically, Zhang et al. (2019b) proposed a new self-distillation method, in which knowledge from the deeper sections of the network is distilled into its shallow sections. Similar to the self-distillation in Zhang et al. (2019b), a self-attention distillation method was proposed

for lane detection (Hou et al. 2019). The network utilizes the attention maps of its own layers as distillation targets for its lower layers. Snapshot distillation (Yang et al. 2019b) is a special variant of self-distillation, in which knowledge in the earlier epochs of the network (teacher) is transferred into its later epochs (student) to support a supervised training process within the same network. To further reduce the inference time via the early exit, Phuong and Lampert (2019b) proposed distillation-based training scheme, in which the early exit layer tries to mimic the output of later exit layer during the training. Recently, self-distillation has been theoretically analyzed in Mobahi et al. (2020), and its improved performance experimentally demonstrated in Zhang and Sabuncu (2020).

Furthermore, some interesting self-distillation methods are recently proposed (Yuan et al. 2020; Yun et al. 2020; Hahn and Choi 2019). To be specific, Yuan *et al.* proposed teacher–free knowledge distillation methods based on the analysis of label smoothing regularization (Yuan et al. 2020). Hahn and Choi proposed a novel self-knowledge distillation method, in which the self-knowledge consists of the predicted probabilities instead of traditional soft probabilities (Hahn and Choi 2019). These predicted probabilities are defined by the feature representations of the training model. They reflect the similarities of data in feature embedding space. Yun *et al.* proposed class-wise self-knowledge distillation to match the output distributions of the training model between intra-class samples and augmented samples within the same source with the same model (Yun et al. 2020). In addition, the self-distillation proposed by Lee et al. (2019a) is adopted for data augmentation and the self-knowledge of augmentation is distilled into the model itself. Self distillation is also adopted to optimize deep models (the teacher or student networks) with the same architecture one by one (Furlanello et al. 2018; Bagherinezhad et al. 2018). Each network distills the knowledge of the previous network using a teacher–student optimization.

Besides, offline, online and self distillation can also be intuitively understood from the perspective of human beings teacher–student learning. Offline distillation means the knowledgeable teacher teaches a student knowledge; online distillation means both teacher and student study together with each other; self-distillation means student learn knowledge by oneself. Moreover, just like the human beings learning, these three kinds of distillation can be combined to complement each other due to their own advantages. For example, both self-distillation and online distillation are properly integrated via the multiple knowledge transfer framework (Sun et al. 2021).
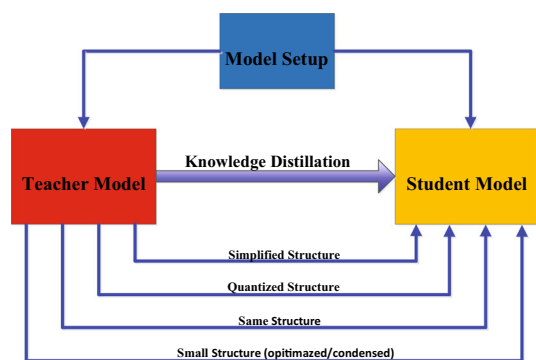
Fig. 9 Relationship of the teacher and student models

4 Teacher–Student Architecture

In knowledge distillation, the teacher–student architecture is a generic carrier to form the knowledge transfer. In other words, the quality of knowledge acquisition and distillation from teacher to student is also determined by how to design the teacher and student networks. In terms of the habits of human beings learning, we hope that a student can find a right teacher. Thus, to well finish capturing and distilling knowledge in knowledge distillation, how to select or design proper structures of teacher and student is very important but difficult problem. Recently, the model setups of teacher and student are almost pre-fixed with unvaried sizes and structures during distillation, so as to easily cause the model capacity gap. However, how to particulary design the architectures of teacher and student and why their architectures are determined by these model setups are nearly missing. In this section, we discuss the relationship between the structures of the teacher model and the student model as illustrated in Fig. 9.

Knowledge distillation was previously designed to compress an ensemble of deep neural networks in Hinton et al. (2015). The complexity of deep neural networks mainly comes from two dimensions: depth and width. It is usually required to transfer knowledge from deeper and wider neural networks to shallower and thinner neural networks (Romero et al. 2015). The student network is usually chosen to be: (1) a simplified version of a teacher network with fewer layers and fewer channels in each layer (Wang et al. 2018a; Zhu and Gong 2018; Li et al. 2020d); or (2) a quantized version of a teacher network in which the structure of the network is preserved (Polino et al. 2018; Mishra and Marr 2018; Wei et al. 2018; Shin et al. 2019); or (3) a small network with efficient basic operations (Howard et al. 2017; Zhang et al. 2018a; Huang et al. 2017); or (4) a small network with optimized global network structure (Liu et al. 2019i; Xie et al. 2020; Gu and Tresp 2020); or (5) the same network as teacher (Zhang et al. 2018b; Furlanello et al. 2018; Tarvainen and Valpola 2017).

The model capacity gap between the large deep neural network and a small student neural network can degrade knowledge transfer (Mirzadeh et al. 2020; Gao et al. 2021). To effectively transfer knowledge to student networks, a variety of methods have been proposed for a controlled reduction of the model complexity (Zhang et al. 2018b; Nowak and Corso 2018; Crowley et al. 2018; Liu et al. 2019a, i; Wang et al. 2018a; Gu and Tresp 2020). Specifically, Mirzadeh et al. (2020) introduced a teacher assistant to mitigate the training gap between teacher model and student model. The gap is further reduced by residual learning, i.e., the assistant structure is used to learn the residual error (Gao et al. 2021). On the other hand, several recent methods also focus on minimizing the difference in structure of the student model and the teacher model. For example, Polino et al. (2018) combined network quantization with knowledge distillation, i.e., the student model is small and quantized version of the teacher model. Nowak and Corso (2018) proposed a structure compression method which involves transferring the knowledge learned by multiple layers to a single layer. Wang et al. (2018a) progressively performed block-wise knowledge transfer from teacher networks to student networks while preserving the receptive field. In online setting, the teacher networks are usually ensembles of student networks, in which the student models share similar structure (or the same structure) with each other (Zhang et al. 2018b; Zhu and Gong 2018; Furlanello et al. 2018; Chen et al. 2020a).

Recently, depth-wise separable convolution has been widely used to design efficient neural networks for mobile or embedded devices (Chollet 2017; Howard et al. 2017; Sandler et al. 2018; Zhang et al. 2018a; Ma et al. 2018). Inspired by the success of neural architecture search (or NAS), the performances of small neural networks have been further improved by searching for a global structure based on efficient meta operations or blocks (Wu et al. 2019; Tan et al. 2019; Tan and Le 2019; Radosavovic et al. 2020). Furthermore, the idea of dynamically searching for a knowledge transfer regime also appears in knowledge distillation, e.g., automatically removing redundant layers in a data-driven way using reinforcement learning (Ashok et al. 2018), and searching for optimal student networks given the teacher networks (Liu et al. 2019i; Xie et al. 2020; Gu and Tresp 2020).

Most previous works focus on designing either the structures of teacher and student models or the knowledge transfer scheme between them. To make a small student model well match a large teacher model for improving knowledge distillation performance, the adaptive teacher–student learning architecture is necessary. Recently, the idea of a neural architecture search in knowledge distillation, i.e., a joint search of student structure and knowledge transfer under the guidance of the teacher model, will be an interesting subject of future study.
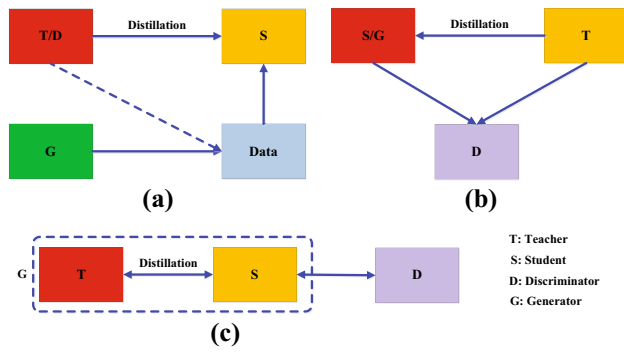
Fig. 10 The different categories of the main adversarial distillation methods. **a** Generator in GAN produces training data to improve KD performance; the teacher may be used as discriminator. **b** Discriminator in GAN ensures that the student (also as generator) mimics the teacher. **c** Teacher and student form a generator; online knowledge distillation is enhanced by the discriminator

5 Distillation Algorithms

A simple yet very effective idea for knowledge transfer is to directly match the response-based knowledge, feature-based knowledge (Romero et al. 2015; Hinton et al. 2015) or the representation distributions in feature space (Passalis and Tefas 2018) between the teacher model and the student model. Many different algorithms have been proposed to improve the process of transferring knowledge in more complex settings. In this section, we review recently proposed typical types of distillation methods for knowledge transfer within the field of knowledge distillation.

5.1 Adversarial Distillation

In knowledge distillation, it is difficult for the teacher model to perfectly learn from the true data distribution. Simultaneously, the student model has only a small capacity and so cannot mimic the teacher model accurately (Mirzadeh et al. 2020). Are there other ways of training the student model in order to mimic the teacher model? Recently, adversarial learning has received a great deal of attention due to its great success in generative networks, i.e., generative adversarial networks or GANs (Goodfellow et al. 2014). Specifically, the discriminator in a GAN estimates the probability that a sample comes from the training data distribution while the generator tries to fool the discriminator using generated data samples. Inspired by this, many adversarial knowledge distillation methods have been proposed to enable the teacher and student networks to have a better understanding of the true data distribution (Wang et al. 2018e; Xu et al. 2018a; Micaelli and Storkey 2019; Xu et al. 2018b; Liu et al. 2018; Wang et al. 2018f; Chen et al. 2019a; Shen et al. 2019d; Shu et al. 2019; Liu et al. 2020a; Belagiannis et al. 2018).

As shown in Fig. 10, adversarial learning-based distillation methods, especially those methods using GANs, can be divided into three main categories as follows. In the first category, an adversarial generator is trained to generate synthetic data, which is either directly used as the training dataset (Chen et al. 2019a; Ye et al. 2020) or used to augment the training dataset (Liu et al. 2018), shown in Fig. 10a. Furthermore, Micaelli and Storkey (2019) utilized an adversarial generator to generate hard examples for knowledge transfer. Generally, the distillation loss used in this GAN-based KD category can be formulated as

$$L_{KD} = \mathcal{L}_G\big(F_t(G(z)), F_s(G(z))\big), \tag{7}$$

where $F_t(.)$ and $F_s(.)$ are the outputs of the teacher and student models, respectively. $G(z)$ indicates the training samples generated by the generator G given the random input vector z, and $\mathcal{L}_G$ is a distillation loss to force the match between the predicted and the ground-truth probability distributions, e.g., the cross entropy loss or the Kullback-Leibler (KL) divergence loss.

To make student well match teacher, a discriminator in the second category is introduced to distinguish the samples from the student and the teacher models by using either the logits (Xu et al. 2018a, b) or the features (Wang et al. 2018f), shown in Fig. 10b. Specifically, Belagiannis et al. (2018) used unlabeled data samples to form the knowledge transfer. Multiple discriminators were used by Shen et al. (2019d). Furthermore, an effective intermediate supervision, i.e., the squeezed knowledge, was used by Shu et al. (2019) to mitigate the capacity gap between the teacher and the student. A representative model proposed by Wang et al. (2018f) falls into this category, which can be formulated as

$$\begin{aligned} L_{GANKD} = \mathcal{L}_{CE}\big(G(F_s(x)), y\big) + \alpha\mathcal{L}_{KL}\big(G(F_s(x)), F_t(x)\big) \\ + \beta\mathcal{L}_{GAN}\big(F_s(x), F_t(x)\big), \end{aligned} \tag{8}$$

where G is a student network and $\mathcal{L}_{GAN}(.)$ indicates a typical loss function used in generative adversarial network to make the outputs between student and teacher as similar as possible.

In the third category, adversarial knowledge distillation is carried out in an online manner, i.e., the teacher and the student are jointly optimized in each iteration (Wang et al. 2018e; Chung et al. 2020), shown in Fig. 10c. Besides, using knowledge distillation to compress GANs, a learned small GAN student network mimics a larger GAN teacher network via knowledge transfer (Aguinaldo et al. 2019; Li et al. 2020c).

In summary, three main points can be concluded from the adversarial distillation methods above as follows: GAN is an effective tool to enhance the power of student learning via the teacher knowledge transfer; joint GAN and KD can generate the valuable data for improving the KD performance
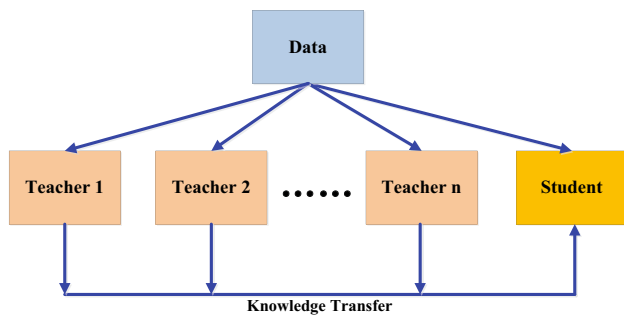
Fig. 11 The generic framework for multi-teacher distillation

Table 3 A summary of multi-teacher distillation using different types of knowledge and distillation schemes. The response-based knowledge, feature-based knowledge and relation-based knowledge are abbreviated as 'ResK', 'FeaK' and 'RelK', respectively

Methods	ResK	FeaK	RelK
Offline Distillation			
You et al. (2017)	✔	✘	✔
Fukuda et al. (2017)	✔	✘	✘
Shen et al. (2019b)	✔	✔	✘
Wu et al. (2019a)	✘	✘	✔
Park and Kwak (2020)	✘	✔	✘
Yang et al. (2020c)	✔	✘	✘
Luo et al. (2020)	✔	✔	✘
Kwon et al. (2020)	✔	✘	✘
Liu et al. (2020c)	✔	✔	✘
Zhao et al. (2020a)	✔	✔	✘
Yuan et al. (2021)	✔	✘	✘
Online Distillation			
Papernot et al. (2017)	✔	✘	✘
Furlanello et al. (2018)	✔	✘	✘
Zhang et al. (2018b)	✔	✘	✘
Yang et al. (2019a)	✔	✘	✘
Lee et al. (2019c)	✔	✔	✘

and overcoming the limitations of unusable and unaccessible data; KD can be used to compress GANs.

5.2 Multi-teacher Distillation

Different teacher architectures can provide their own useful knowledge for a student network. The multiple teacher networks can be individually and integrally used for distillation during the period of training a student network. In a typical teacher–student framework, the teacher usually has a large model or an ensemble of large models. To transfer knowledge from multiple teachers, the simplest way is to use the averaged response from all teachers as the supervision signal (Hinton et al. 2015). Several multi-teacher knowledge distillation methods have recently been proposed (Sau and Balasubramanian 2016; You et al. 2017; Chen et al. 2019b; Furlanello et al. 2018; Yang et al. 2019a; Zhang et al. 2018b; Lee et al. 2019c; Park and Kwak 2020; Papernot et al. 2017; Fukuda et al. 2017; Ruder et al. 2017; Wu et al. 2019a; Yang et al. 2020c; Vongkulbhisal et al. 2019; Zhao et al. 2020a; Yuan et al. 2021). A generic framework for multi-teacher distillation is shown in Fig. 11.

Multiple teacher networks have turned out to be effective for training student model usually using logits and feature representation as the knowledge. In addition to the averaged logits from all teachers, You et al. (2017) further incorporated features from the intermediate layers in order to encourage the dissimilarity among different training samples. To utilize both logits and intermediate features, Chen et al. (2019b) used two teacher networks, in which one teacher transfers response-based knowledge to the student and the other teacher transfers feature-based knowledge to the student. Fukuda et al. (2017) randomly selected one teacher from the pool of teacher networks at each iteration. To transfer feature-based knowledge from multiple teachers, additional teacher branches are added to the student networks to mimic the intermediate features of teachers (Park and Kwak 2020; Asif et al. 2020). Born again networks address multiple teachers in a step-by-step manner, i.e., the student at the t step is used as the teacher of the student at the $t + 1$ step (Furlanello et al. 2018),

and similar ideas can be found in Yang et al. (2019a). To efficiently perform knowledge transfer and explore the power of multiple teachers, several alternative methods have been proposed to simulate multiple teachers by adding different types of noise to a given teacher (Sau and Balasubramanian 2016) or by using stochastic blocks and skip connections (Lee et al. 2019c). Using multiple teacher models with feature ensembles, knowledge amalgamation is designed in (Shen et al. 2019a; Luo et al. 2019; Shen et al. 2019b; Luo et al. 2020). Through knowledge amalgamation, many public available trained deep models as teachers can be reused. More interestingly, due to the special characteristics of multi-teacher distillation, its extensions are used for domain adaptation via knowledge adaptation (Ruder et al. 2017), and to protect the privacy and security of data (Vongkulbhisal et al. 2019; Papernot et al. 2017).

A summary of typical multi-teacher distillation methods using different types of knowledge and distillation schemes is shown in Table 3. Generally, multi-teacher knowledge distillation can provide rich knowledge and tailor a versatile student model because of the diverse knowledge from different teachers. However, how to effectively integrate different types of knowledge from multiple teachers needs to be further studied.
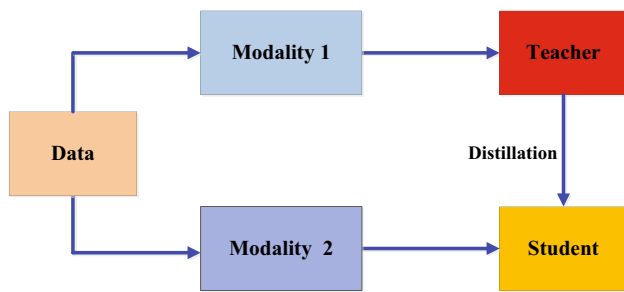
Fig. 12 The generic framework for cross-modal distillation. For simplicity, only two modalities are shown

5.3 Cross-Modal Distillation

The data or labels for some modalities might not be available during training or testing (Gupta et al. 2016; Garcia et al. 2018; Zhao et al. 2018; Roheda et al. 2018; Zhao et al. 2020b). For this reason it is important to transfer knowledge between different modalities. Several typical scenarios using cross-modal knowledge transfer are reviewed as follows.

Given a teacher model pretrained on one modality (e.g., RGB images) with a large number of well-annotated data samples, Gupta et al. (2016) transferred the knowledge from the teacher model to the student model with a new unlabeled input modality, such as a depth image and optical flow. Specifically, the proposed method relies on unlabeled paired samples involving both modalities, i.e., both RGB and depth images. The features obtained from RGB images by the teacher are then used for the supervised training of the student (Gupta et al. 2016). The idea behind the paired samples is to transfer the annotation or label information via pair-wise sample registration and has been widely used for cross-modal applications (Albanie et al. 2018; Zhao et al. 2018; Thoker and Gall 2019). To perform human pose estimation through walls or with occluded images, Zhao et al. (2018) used synchronized radio signals and camera images. Knowledge is transferred across modalities for radio-based human pose estimation. Thoker and Gall (2019) obtained paired samples from two modalities: RGB videos and skeleton sequence. The pairs are used to transfer the knowledge learned on RGB videos to a skeleton-based human action recognition model. To improve the action recognition performance using only RGB images, Garcia et al. (2018) performed cross-modality distillation on an additional modality, i.e., depth image, to generate a hallucination stream for RGB image modality. Tian et al. (2020) introduced a contrastive loss to transfer pair-wise relationship across different modalities. To improve target detection, Roheda et al. (2018) proposed cross-modality distillation among the missing and available modelities using GANs. The generic framework of cross-modal distillation is shown in Fig. 12.

Moreover, Do et al. (2019) proposed a knowledge distillation-based visual question answering method, in which knowledge from trilinear interaction teacher model with image-question-answer as inputs is distilled into the learning of a bilinear interaction student model with image-question as inputs. The probabilistic knowledge distillation proposed by Passalis and Tefas (2018) is also used for knowledge transfer from the textual modality into the visual modality. Hoffman et al. (2016) proposed a modality hallucination architecture based on cross-modality distillation to improve detection performance. Besides, these cross-model distillation methods also transfer the knowledge among multiple domains (Kundu et al. 2019; Chen et al. 2019c; Su and Maji 2017).

A summary of cross-modal distillation with different modalities, types of knowledge and distillation schemes is shown in Table 4. Specifically, it can be seen that knowledge distillation performs well in visual recognition tasks in the cross-modal scenarios. However, cross-modal knowledge transfer is a challenging study when there is a modality gap, e.g., lacking of the paired samples between different modalities.
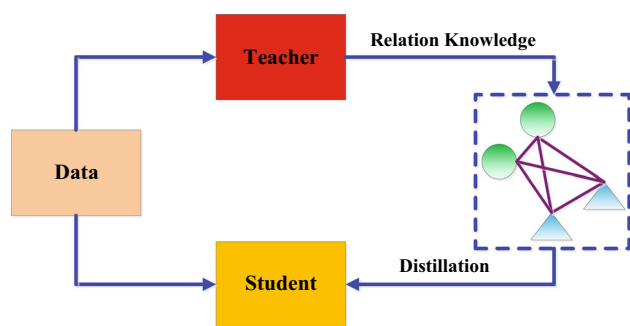
5.4 Graph-Based Distillation

Most of knowledge distillation algorithms focus on transferring individual instance knowledge from the teacher to the student, while some recent methods have been proposed to explore the intra-data relationships using graphs (Chen et al. 2021; Zhang and Peng 2018; Lee and Song 2019; Park et al. 2019; Yao et al. 2020; Ma and Mei 2019; Hou et al. 2020). The main ideas of these graph-based distillation methods are 1) to use the graph as the carrier of teacher knowledge; or 2) to use the graph to control the message passing of the teacher knowledge. A generic framework for graph-based distillation is shown in Fig. 13. As described in Sect. 2.3, the graph-based knowledge falls in line of relation-based knowledge. In this section, we introduce typical definitions of the graph-based knowledge and the graph-based message passing distillation algorithms.

Specifically, in Zhang and Peng (2018), each vertex represents a self-supervised teacher. Two graphs are then constructed using logits and intermediate features, i.e., the logits graph and representation graph, to transfer knowledge from multiple self-supervised teachers to the student. In Chen et al. (2021), the graph is used to maintain the relationship between samples in the high-dimensional space. Knowledge transfer is then carried out using a proposed locality preserving loss function. Lee and Song (2019) analysed intra-data relations using a multi-head graph, in which the vertices are the features from different layers in CNNs. Park et al. (2019) directly transferred the mutual relations of data samples, i.e., to match edges between a teacher graph and a student graph. Tung and Mori (2019) used the similarity matrix to represent

Table 4 A summary of cross-modal distillation with modalities, types of knowledge and distillation

Methods	Modality for teacher	Modality for student	Knowledge	Distillation
Hoffman et al. (2016)	RGB images	Depth images	FeaK	Offline
Gupta et al. (2016)	RGB images	Depth images	ResK	Offline
Passalis and Tefas (2018)	Textual modality	Visual modality	RelK	Offline
Garcia et al. (2018)	Depth and RGB videos	RGB videos	ResK, FeaK	Offline
Zhao et al. (2018)	RGB frames	Radio frequency heatmaps	ResK	Offline
Roheda et al. (2018)	Temporal data	Spatial data	FeaK	Online
Albanie et al. (2018)	Vision	Sound	ResK	Offline
Thoker and Gall (2019)	RGB videos	Skeleton data	ResK	Offline
Do et al. (2019)	Images, question, answer information	Image-questions	ResK	Offline
Tian et al. (2020)	RGB images	Depth images	ResK	Offline
Gao et al. (2020)	Multi-modal images	Single-mode images	ResK, FeaK	Offline

Fig. 13 A generic framework for graph-based distillation

the mutual relations of the activations of the input pairs in teacher and student models. The similarity matrix of student matches that of teacher. Furthermore, Peng et al. (2019a) not only matched the response-based and feature-based knowledge, but also used the graph-based knowledge. In Liu et al. (2019g), the instance features and instance relationships are modeled as vertexes and edges of the graph, respectively.

Rather than using the graph-based knowledge, several methods control knowledge transfer using a graph. Specifically, Luo et al. (2018) considered the modality discrepancy to incorporate privileged information from the source domain. A directed graph, referred to as a distillation graph is introduced to explore the relationship between different modalities. Each vertex represent a modality and the edges indicate the connection strength between one modality and another. Minami et al. (2019) proposed a bidirectional graph-based diverse collaborative learning to explore diverse knowledge transfer patterns. Yao et al. (2020) introduced GNNs to deal with the knowledge transfer for graph-based knowledge. Besides, using knowledge distillation, the topological semantics of a graph convolutional teacher network as the topology-aware knowledge are transferred into the graph convolutional student network (Yang et al. 2020b).

Graph-based distillation can transfer the informative structure knowledge of data. However, how to properly construct graph to model the structure knowledge of data is a still challenging study.

5.5 Attention-Based Distillation

Since attention can well reflect the neuron activations of convolutional neural network, some attention mechanisms are used in knowledge distillation to improve the performance of the student network (Zagoruyko and Komodakis 2017; Huang and Wang 2017; Srinivas and Fleuret 2018; Crowley et al. 2018; Song et al. 2018). Among these attention-based KD methods (Crowley et al. 2018; Huang and Wang 2017; Srinivas and Fleuret 2018; Zagoruyko and Komodakis 2017), different attention transfer mechanisms are defined for distilling knowledge from the teacher network to the student network. The core of attention transfer is to define the attention maps for feature embedding in the layers of a neural network. That is to say, knowledge about feature embedding is transferred using attention map functions. Unlike the attention maps, a different attentive knowledge distillation method was proposed by Song et al. (2018). An attention mechanism is used to assign different confidence rules (Song et al. 2018).

5.6 Data-Free Distillation

Some data-free KD methods have been proposed to overcome problems with unavailable data arising from privacy, legality, security and confidentiality concerns (Chen et al. 2019a; Lopes et al. 2017; Nayak et al. 2019; Micaelli and Storkey 2019; Haroush et al. 2020; Ye et al. 2020; Nayak et al. 2021; Chawla et al. 2021). Just as "data free" implies, there is no training data. Instead, the data is newly or synthetically generated.
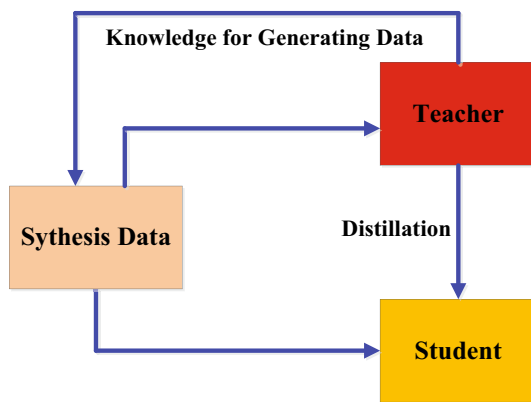
Fig. 14 A generic framework for data-free distillation

Fig. 15 A generic framework for quantized distillation

Specifically, in (Chen et al. 2019a; Ye et al. 2020; Micaelli and Storkey 2019; Yoo et al. 2019; Hu et al. 2020), the transfer data is generated by a GAN. In the proposed data-free knowledge distillation method (Lopes et al. 2017), the transfer data to train the student network is reconstructed by using the layer activations or layer spectral activations of the teacher network. Yin et al. (2020) proposed DeepInversion, which uses knowledge distillation to generate synthesized images for data-free knowledge transfer. Nayak et al. (2019) proposed zero-shot knowledge distillation that does not use existing data. The transfer data is produced by modelling the softmax space using the parameters of the teacher network. In fact, the target data in (Micaelli and Storkey 2019; Nayak et al. 2019) is generated by using the information from the feature representations of teacher networks. Similar to zero-shot learning, a knowledge distillation method with few-shot learning is designed by distilling knowledge from a teacher model into a student neural network (Kimura et al. 2018; Shen et al. 2021). The teacher uses limited labelled data. Besides, there is a new type of distillation called data distillation, which is similar to data-free distillation (Radosavovic et al. 2018; Liu et al. 2019d; Zhang et al. 2020d). In data distillation, new training annotations of unlabeled data generated from the teacher model are employed to train a student model.

In summary, the synthesis data in data-free distillation is usually generated from the feature representations from the pre-trained teacher model, as shown in Fig. 14. Although the data-free distillation has shown a great potential under the condition of unavailable data, it remains a very challenging task, i.e., how to generate high quality diverse training data to improve the model generalizability.

5.7 Quantized Distillation

Network quantization reduces the computation complexity of neural networks by converting high-precision networks (e.g.,
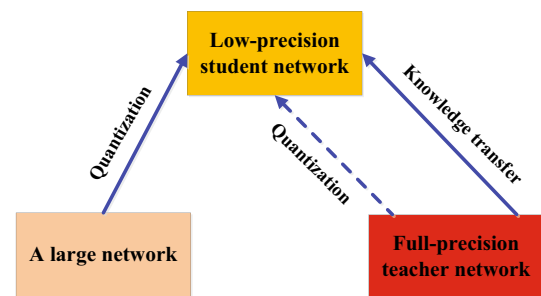
32-bit floating point) into low-precision networks (e.g., 2-bit and 8-bit). Meanwhile, knowledge distillation aims to train a small model to yield a performance comparable to that of a complex model. Some KD methods have been proposed using the quantization process in the teacher–student framework (Polino et al. 2018; Mishra and Marr 2018; Wei et al. 2018; Shin et al. 2019; Kim et al. 2019a). A framework for quantized distillation methods is shown in Fig. 15.

Specifically, Polino et al. (2018) proposed a quantized distillation method to transfer the knowledge to a weight-quantized student network. In Mishra and Marr (2018), the proposed quantized KD is called the "apprentice" . A high precision teacher network transfers knowledge to a small low-precision student network. To ensure that a small student network accurately mimics a large teacher network, the full-precision teacher network is first quantized on the feature maps, and then the knowledge is transferred from the quantized teacher to a quantized student network (Wei et al. 2018). Kim et al. (2019a) proposed quantization-aware knowledge distillation, which is based on self-study of a quantized student network and on the co-studying of teacher and student networks with knowledge transfer. Furthermore, Shin et al. (2019) carried out empirical analysis of deep neural networks using both distillation and quantization, taking into account the hyper-parameters for knowledge distillation, such as the size of teacher networks and the distillation temperature. Recently, unlike the quantized distillation methods above, a self-distillation training schemes is designed to improve the performance of quantized deep models, where teacher shares model parameters of student (Boo et al. 2021).

5.8 Lifelong Distillation

Lifelong learning, including continual learning, continuous learning and meta-learning, aims to learn in a similar way to human. It accumulates the previously learned knowledge and also transfers the learned knowledge into future learning (Chen and Liu 2018). Knowledge distillation provides an effective way to preserve and transfer learned knowledge without catastrophic forgetting. Recently, an increasing number of KD variants, which are based on lifelong learning, have

been developed (Jang et al. 2019; Flennerhag et al. 2019; Peng et al. 2019b; Liu et al. 2019e; Lee et al. 2019b; Zhai et al. 2019; Zhou et al. 2020; Shmelkov et al. 2017; Li and Hoiem 2017; Caccia et al. 2020). The methods proposed in (Jang et al. 2019; Peng et al. 2019b; Liu et al. 2019e; Flennerhag et al. 2019) adopt meta-learning. Jang et al. (2019) designed meta-transfer networks that can determine what and where to transfer in the teacher–student architecture. Flennerhag et al. (2019) proposed a light-weight framework called Leap for meta-learning over task manifolds by transferring knowledge from one learning process to another. Peng et al. (2019b) designed a new knowledge transfer network architecture for few-shot image recognition. The architecture simultaneously incorporates visual information from images and prior knowledge. Liu et al. (2019e) proposed the semantic-aware knowledge preservation method for image retrieval. The teacher knowledge obtained from the image modalities and semantic information are preserved and transferred.

Moreover, to address the problem of catastrophic forgetting in lifelong learning, global distillation (Lee et al. 2019b), knowledge distillation-based lifelong GAN (Zhai et al. 2019), multi-model distillation (Zhou et al. 2020) and the other KD-based methods (Li and Hoiem 2017; Shmelkov et al. 2017) have been developed to extract the learned knowledge and teach the student network on new tasks.

5.9 NAS-Based Distillation

Neural architecture search (NAS), which is one of the most popular auto machine learning (or AutoML) techniques, aims to automatically identify deep neural models and adaptively learn appropriate deep neural structures. In knowledge distillation, the success of knowledge transfer depends on not only the knowledge from the teacher but also the architecture of the student. However, there might be a capacity gap between the large teacher model and the small student model, making it difficult for the student to learn well from the teacher. To address this issue, neural architecture search has been adopted to find the appropriate student architecture in oracle-based (Kang et al. 2020) and architecture-aware knowledge distillation (Liu et al. 2019i). Furthermore, knowledge distillation is employed to improve the efficiency of neural architecture search, such as AdaNAS (Macko et al. 2019), NAS with distilled architecture knowledge (Li et al. 2020a), teacher guided search for architectures or TGSA (Bashivan et al. 2019), and one-shot NAS (Peng et al. 2020). In TGSA, each architecture search step is guided to mimic the intermediate feature representations of the teacher network. The possible structures for the student are efficiently searched and the feature transfer is effectively supervised by the teacher.

6 Performance Comparison

Knowledge distillation is an excellent technique for model compression. Through capturing the teacher knowledge and using distillation strategies with teacher–student learning, it provides effective performance of the lightweight student model. Recently, many knowledge distillation methods focus on improving the performance, especially in image classification tasks. In this section, to clearly demonstrate the effectiveness of knowledge distillation, we summarize the classification performance of some typical KD methods on two popular image classification datasets.

The two datasets are CIFAR10 and CIFAR100 (Krizhevsky and Hinton 2009) that are composed of 32×32 RGB images taken from 10 and 100 classes, respectively. Both have 50000 training images and 10000 testing images, and each class has the same numbers of training and testing images. For fair comparison, the experimental classification accuracy results (%) of the KD methods are directly derived from the corresponding original papers, as shown in Table 5 for CIFAR10 and Table 6 for CIFAR100. We report the performance of different methods when using different types of knowledge, distillation schemes, and structures of teacher/student models. Specifically, the accuracies in parentheses are the classification results of the teacher and student models, which are trained individually. It should be noted that the pairs of accuracies of DML (Zhang et al. 2018b), DCM (Yao and Sun 2020) and KDCL (Guo et al. 2020) are the performance of teacher and student after online distillation.

From the performance comparison in Table 5 and Table 6, several observations can be summarized as

- Knowledge distillation can be realized on different deep models.
- Model compression of different deep models can be achieved by knowledge distillation.
- The online knowledge distillation through collaborative learning (Zhang et al. 2018b; Yao and Sun 2020) can significantly improve the performance of the deep models.
- The self-knowledge distillation (Yang et al. 2019b; Yuan et al. 2020; Xu and Liu 2019; Yun et al. 2020) can well improve the performance of the deep models.
- The offline and online distillation methods often transfer feature-based knowledge and response-based knowledge, respectively.
- The performance of the lightweight deep models (student) can be improved by the knowledge transfer from the high-capacity teacher models.

Through the performance comparison of different knowledge distillation methods, it can be easily concluded that knowledge distillation is an effective and efficient technique of compressing deep models.

Table 5 Performance comparison of different knowledge distillation methods on CIFAR10. Note that ↑ indicates the performance improvement of the student network learned by each method comparing with the corresponding baseline model

Methods	Knowledge	Teacher (baseline)	Student (baseline)	Accuracies
Offline distillation				
FSP (Yim et al. 2017)	RelK	ResNet26 (91.91)	ResNet8 (87.91)	88.70 (0.79 ↑)
FT (Kim et al. 2018)	FeaK	ResNet56 (93.61)	ResNet20 (92.22)	93.15 (0.93 ↑)
IRG (Liu et al. 2019g)	RelK	ResNet20 (91.45)	ResNet20-x0.5 (88.36)	90.69 (2.33 ↑)
SP (Tung and Mori 2019)	RelK	WRN-40-1 (93.49)	WRN-16-1 (91.26)	91.87 (0.61 ↑)
SP (Tung and Mori 2019)	RelK	WRN-40-2 (95.76)	WRN-16-8 (94.82)	95.45 (0.63 ↑)
FN (Xu et al. 2020b)	FeaK	ResNet110 (94.29)	ResNet56 (93.63)	94.14 (0.51 ↑)
FN (Xu et al. 2020b)	FeaK	ResNet56 (93.63)	ResNet20 (92.11)	92.67 (0.56 ↑)
AdaIN (Yang et al. 2020a)	FeaK	ResNet26 (93.58)	ResNet8 (87.78)	89.02 (1.24 ↑)
AdaIN (Yang et al. 2020a)	FeaK	WRN-40-2 (95.07)	WRN-16-2 (93.98)	94.67 (0.69 ↑)
AE-KD (Du et al. 2020)	FeaK	ResNet56 (–)	MobileNetV2 (75.97)	77.07 (1.10 ↑)
JointRD (Li et al. 2020b)	FeaK	ResNet34 (95.39)	plain-CNN 34 (93.73)	94.78 (1.05 ↑)
TOFD (Zhang et al. 2020a)	FeaK	ResNet152 (–)	ResNeXt50-4 (94.49)	97.09 (2.60 ↑)
TOFD (Zhang et al. 2020a)	FeaK	ResNet152 (–)	MobileNetV2 (90.43)	93.34 (2.91 ↑)
CTKD (Zhao et al. 2020a)	RelK, FeaK	WRN-40-1 (93.43)	WRN-16-1 (91.28)	92.50 (1.22 ↑)
CTKD (Zhao et al. 2020a)	RelK, FeaK	WRN-40-2 (94.70)	WRN-16-2 (93.68)	94.42 (0.74 ↑)
Online distillation				
Rocket-KD (Zhou et al. 2018)	FeaK	WRN-40-1 (93.42)	WRN-16-1 (91.23)	92.48 (1.25 ↑)
DML (Zhang et al. 2018b)	ResK	WRN-28-10 (95.01)	ResNet32 (92.47)	95.75, 93.18 (0.71 ↑)
DML (Zhang et al. 2018b)	ResK	MobileNet (93.59)	ResNet32 (92.47)	94.24, 93.32 (0.85 ↑)
DML (Zhang et al. 2018b)	ResK	ResNet32 (92.47)	ResNet32 (92.47)	92.68, 92.80 (0.33 ↑)
ONE (Zhu and Gong 2018)	ResK	ResNet32+ONE	ResNet32 (93.07)	94.01 (0.84 ↑)
ONE (Zhu and Gong 2018)	ResK	ResNet110+ONE	ResNet110 (94.44)	94.83 (0.39 ↑)
PCL (Wu and Gong 2021)	ResK	Student ensemble	ResNet110 (94.91)	95.53 (0.62 ↑)
PCL (Wu and Gong 2021)	ResK	Student ensemble	DenseNet-40-12 (93.19)	94.13 (0.94 ↑)
PCL (Wu and Gong 2021)	ResK	Student ensemble	VGG16 (93.96)	94.74 (0.78 ↑)
ACNs (Zhang et al. 2021a)	ResK	ResNet14 (90.66)	ResNet14 (90.66)	92.09 (1.43 ↑)
ACNs (Zhang et al. 2021a)	ResK	VGG11 (91.25)	VGG11 (91.25)	92.65 (1.40 ↑)
ACNs (Zhang et al. 2021a)	ResK	AlexNet (73.24)	AlexNet (73.24)	78.57 (5.33 ↑)
Self-Distillation				
Xu and Liu (2019)	ResK, FeaK	–	ResNet32 (92.78)	93.68 (0.90↑)
Xu and Liu (2019)	ResK, FeaK	–	DenseNe40(94.53)	94.80 (0.27↑)

7 Applications

As an effective technique for the compression and acceleration of deep neural networks, knowledge distillation has been widely used in different fields of artificial intelligence, including visual recognition, speech recognition, natural language processing (NLP), and recommendation systems. Furthermore, knowledge distillation also can be used for other purposes, such as the data privacy and as a defense against adversarial attacks. This section briefly reviews applications of knowledge distillation.

7.1 KD in Visual Recognition

In last few years, a variety of knowledge distillation methods have been widely used for model compression in different visual recognition applications. Specifically, most of the knowledge distillation methods were previously developed for image classification (Li and Hoiem 2017; Peng et al. 2019b; Bagherinezhad et al. 2018; Chen et al. 2018a; Wang et al. 2019b; Mukherjee et al. 2019; Zhu et al. 2019) and then extended to other visual recognition applications, including face recognition (Luo et al. 2016; Kong et al. 2019; Yan et al. 2019; Ge et al. 2018; Wang et al. 2018b, 2019c; Duong et al. 2019; Wu et al. 2020; Wang et al. 2017; Zhang et al.

Table 6 Performance comparison of different knowledge distillation methods on CIFAR100. Note that ↑ indicates the performance improvement of the student network learned by each method comparing with the corresponding baseline model

Methods	Knowledge	Teacher (baseline)	Student (baseline)	Accuracies
Offline distillation				
FSP (Yim et al. 2017)	RelK	ResNet32 (64.06)	ResNet14 (58.65)	63.33 (4.68 ↑)
FT (Kim et al. 2018)	FeaK	ResNet110 (73.09)	ResNet56 (71.96)	74.48 (2.52 ↑)
RKD (Park et al. 2019)	RelK, FeaK	ResNet50 (77.76)	VGG11 (71.26)	74.66 (3.40 ↑)
IRG (Liu et al. 2019g)	RelK	ResNet20 (78.40)	ResNet20-x0.5 (72.51)	74.64 (2.13 ↑)
CCKD (Peng et al. 2019a)	RelK, ResK	ResNet110 (–)	ResNet20 (68.40)	72.40 (4.00 ↑)
KR (Liu et al. 2019c)	FeaK	ResNet32 (64.06)	ResNet14 (58,65)	63.95 (5.30 ↑)
LKD (Li et al. 2020e)	RelK	ResNet110 (75.76)	ResNet20 (69.47)	72.63 (3.16 ↑)
LKD (Li et al. 2020e)	RelK	WRN-40-2 (75.61)	WRN-16-2 (73.10)	75.44 (2.34 ↑)
SSKD (Xu et al. 2020a)	RelK, ResK	VGG13 (75.38)	MobileNetV2 (65.79)	71.53 (5.74 ↑)
SSKD (Xu et al. 2020a)	RelK, ResK	ResNet50 (79.10)	MobileNetV2 (65.79)	72.57 (6.78 ↑)
FN (Xu et al. 2020b)	FeaK	ResNet110 (82.01)	ResNet56 (81.73)	82.23 (0.50 ↑)
AdaIN (Yang et al. 2020a)	FeaK	WRN-40-4 (78.31)	WRN-16-4 (75.68)	78.25 (2.57 ↑)
AdaIN (Yang et al. 2020a)	FeaK	ResNet34 (77.26)	MobileNetV2 (68.36)	70.66 (2.30 ↑)
PAD-L_2 (Zhang et al. 2020e)	FeaK	ResNet18 (75.86)	MobileNetV2 (68.16)	74.06 (5.90 ↑)
MGD (Yue et al. 2020)	FeaK	WRN-28-4 (78.91)	WRN-28-2 (75.12)	78.82 (3.70 ↑)
AE-KD (Du et al. 2020)	FeaK	ResNet56 (–)	ResNet20 (69.06)	70.55 (1.49 ↑)
JointRD (Li et al. 2020b)	FeaK	ResNet18 (77.92)	plain-CNN 18 (77.44)	78.24 (0.80 ↑)
TOFD (Zhang et al. 2020a)	FeaK	ResNet152 (–)	ResNet50 (77.42)	84.74 (7.32 ↑)
TOFD (Zhang et al. 2020a)	FeaK	ResNet152 (–)	ShuffleNetV2 (72.38)	76.68 (4.30 ↑)
CTKD (Zhao et al. 2020a)	RelK, FeaK	ResNet110 (72.65)	ResNet20 (68.33)	70.75 (2.42 ↑)
CTKD (Zhao et al. 2020a)	RelK, FeaK	WRN-40-2 (75.42)	WRN-16-2 (72.27)	74.70 (2.43 ↑)
SemCKD (Chen et al. 2021)	FeaK	ResNet-32x4 (79.42)	VGG13 (74.82)	79.43 (4.61 ↑)
SemCKD (Chen et al. 2021)	FeaK	WRN-40-2 (75.61)	MobileNetV2 (65.43)	69.61 (4.18 ↑)
SemCKD (Chen et al. 2021)	FeaK	VGG13 (74.64)	ShuffleNetV2 (72.60)	76.39 (3.79 ↑)
RKD (Gao et al. 2021)	FeaK	ResNet34 (73.05)	ResNet18 (68.06)	72.82 (4.76 ↑)
Online distillation				
Rocket-KD (Zhou et al. 2018)	FeaK	WRN-40-1 (–)	WRN-16-1 (56.30)	67.00 (10.07 ↑)
DML (Zhang et al. 2018b)	ResK	WRN-28-10 (78.69)	MobileNet (73.65)	80.28, 77.39 (3.74 ↑)
DML (Zhang et al. 2018b)	ResK	MobileNet (73.65)	ResNet32 (68.99)	76.13, 71.10 (8.11 ↑)
ONE (Zhu and Gong 2018)	ResK	ResNet32+ONE	ResNet32 (68.82)	73.39 (4.57 ↑)
ONE (Zhu and Gong 2018)	ResK	ResNet110+ONE	ResNet110 (74.67)	78.38 (3.71 ↑)
DCM (Yao and Sun 2020)	ResK	WRN-28-10 (81.28)	ResNet110 (73.45)	82.18, 77.01 (3.56 ↑)
DCM (Yao and Sun 2020)	ResK	WRN-28-10 (81.28)	MobileNet (73.70)	83.17, 78.57 (4.87 ↑)
KDCL (Guo et al. 2020)	ResK	WRN-16-2 (72.20)	ResNet32 (69.90)	75.50, 74.30 (4.40 ↑)
PCL (Wu and Gong 2021)	ResK	Student ensemble	ResNet110 (76.21)	79.98 (3.77 ↑)
PCL (Wu and Gong 2021)	ResK	Student ensemble	DenseNet-40-12 (71.03)	73.09 (2.06 ↑)
ACNs (Zhang et al. 2021a)	ResK	ResNet14 (66.88)	ResNet14 (66.88)	68.40 (1.52 ↑)
ACNs (Zhang et al. 2021a)	ResK	VGG11 (67.38)	VGG11 (67.38)	70.11 (2.73 ↑)
ACNs (Zhang et al. 2021a)	ResK	AlexNet (39.45)	AlexNet (39.45)	46.27 (6.82 ↑)
Self-distillation				
Xu and Liu (2019)	ResK, FeaK	–	DenseNet (74.80)	76.32 (1.52↑)
SD (Yang et al. 2019b)	ResK	–	ResNet32 (68.39)	71.29 (2.90↑)
Tf-KD (Yuan et al. 2020)	ResK	–	ResNet18 (75.87)	77.10 (1.23↑)
Tf-KD (Yuan et al. 2020)	ResK	–	ShuffleNetV2 (70.34)	72.23 (1.89↑)
Tf-KD (Yuan et al. 2020)	ResK	–	ResNeXt29 (81.03)	82.08 (1.05↑)
CS-KD (Yun et al. 2020)	ResK	–	ResNet18 (75.29)	78.01 (2.72↑)

2020b; Wang et al. 2020b), image/video segmentation (He et al. 2019; Mullapudi et al. 2019; Dou et al. 2020; Liu et al. 2019h; Siam et al. 2019; Hou et al. 2020; Bergmann et al. 2020), action recognition (Luo et al. 2018; Hao and Zhang 2019; Thoker and Gall 2019; Garcia et al. 2018; Wang et al. 2019e; Wu et al. 2019b; Zhang et al. 2020c; Cui et al. 2020), object detection (Li et al. 2017; Shmelkov et al. 2017; Cun and Pun 2020; Wang et al. 2019d; Huang et al. 2020; Wei et al. 2018; Hong and Yu 2019; Chawla et al. 2021), lane detection (Hou et al. 2019), person re-identification (Wu et al. 2019a), pedestrian detection (Shen et al. 2016), facial land-mark detection (Dong and Yang 2019), pose estimation (Nie et al. 2019; Zhang et al. 2019a; Zhao et al. 2018), video cap-tioning (Pan et al. 2020; Zhang et al. 2020f), person search (Munjal et al. 2019; Zhang et al. 2021c), image retrieval (Liu et al. 2019e), shadow detection (Chen et al. 2020c), saliency estimation (Li et al. 2019), depth estimation (Pilzer et al. 2019; Ye et al. 2019), visual odometry (Saputra et al. 2019), text-to-image synthesis (Yuan and Peng 2020; Tan et al. 2021), video classification (Zhang and Peng 2018; Bhard-waj et al. 2019), visual question answering (Mun et al. 2018; Aditya et al. 2019) and anomaly detection (Bergmann et al. 2020). Since knowledge distillation in classification task is fundamental for other tasks, we briefly review knowledge distillation in challenging image classification settings, such as face recognition and action recognition.

Existing KD-based face recognition methods focus on not only efficient deployment but also competitive recognition accuracy (Luo et al. 2016; Kong et al. 2019; Yan et al. 2019; Ge et al. 2018; Wang et al. 2018b, 2019c; Duong et al. 2019; Wang et al. 2017, 2020b; Zhang et al. 2020b). Specifically, in Luo et al. (2016), the knowledge from the chosen infor-mative neurons of top hint layer of the teacher network is transferred into the student network. A teacher weighting strategy with the loss of feature representations from hint lay-ers was designed for knowledge transfer to avoid the incorrect supervision by the teacher (Wang et al. 2018b). A recur-sive knowledge distillation method was designed by using a previous student network to initialize the next one (Yan et al. 2019). Since most face recognition methods perform the open-set recognition, i.e., the classes/identities on test set are unknown to the training set, the face recognition criteria are usually distance metrics between feature representations of positive and negtive samples, e.g., the angular loss in Duong et al. (2019) and the correlated embedding loss in Wu et al. (2020).

To improve low-resolution face recognition accuracy, the knowledge distillation framework is developed by using architectures between high-resolution face teacher and low-resolution face student for model acceleration and improved classification performance (Ge et al. 2018; Wang et al. 2019c; Kong et al. 2019; Ge et al. 2020). Specifically, Ge et al. (2018) proposed a selective knowledge distillation method,

in which the teacher network for high-resolution face recog-nition selectively transfers its informative facial features into the student network for low-resolution face recognition through sparse graph optimization. In Kong et al. (2019), cross-resolution face recognition was realized by designing a resolution invariant model unifying both face hallucination and heterogeneous recognition sub-nets. To get efficient and effective low resolution face recognition model, the multi-kernel maximum mean discrepancy between student and teacher networks was adopted as the feature loss (Wang et al. 2019c). In addition, the KD-based face recognition can be extended to face alignment and verification by changing the losses in knowledge distillation (Wang et al. 2017).

Recently, knowledge distillation has been used success-fully for solving the complex image classification problems (Zhu et al. 2019; Bagherinezhad et al. 2018; Peng et al. 2019b; Li and Hoiem 2017; Chen et al. 2018a; Wang et al. 2019b; Mukherjee et al. 2019). For incomplete, ambiguous and redundant image labels, the label refinery model through self-distillation and label progression is proposed to learn soft, informative, collective and dynamic labels for complex image classification (Bagherinezhad et al. 2018). To address catastrophic forgetting with CNN in a variety of image classi-fication tasks, a learning without forgetting method for CNN, including both knowledge distillation and lifelong learning is proposed to recognize a new image task and to preserve the original tasks (Li and Hoiem 2017). For improving image classification accuracy, Chen et al. (2018a) proposed the fea-ture maps-based knowledge distillation method with GAN. It transfers knowledge from feature maps to a student. Using knowledge distillation, a visual interpretation and diagno-sis framework that unifies the teacher–student models for interpretation and a deep generative model for diagnosis is designed for image classifiers (Wang et al. 2019b). Simi-lar to the KD-based low-resolution face recognition, Zhu et al. (2019) proposed deep feature distillation for the low-resolution image classification, in which the output features of a student match that of teacher.

As argued in Sect. 5.3, knowledge distillation with the teacher–student structure can transfer and preserve the cross-modality knowledge. Efficient and effective action recogni-tion under its cross-modal task scenarios can be successfully realized (Thoker and Gall 2019; Luo et al. 2018; Gar-cia et al. 2018; Hao and Zhang 2019; Wu et al. 2019b; Zhang et al. 2020c). These methods are the examples of spa-tiotemporal modality distillation with a different knowledge transfer for action recognition. Examples include mutual teacher–student networks (Thoker and Gall 2019), multiple stream networks (Garcia et al. 2018), spatiotemporal dis-tilled dense-connectivity network (Hao and Zhang 2019), graph distillation (Luo et al. 2018) and multi-teacher to multi-student networks (Wu et al. 2019b; Zhang et al. 2020c). Among these methods, the light-weight student can distill

and share the knowledge information from multiple modalities stored in the teacher.

We summarize two main observations of distillation-based visual recognition applications, as follows.

- Knowledge distillation provides efficient and effective teacher–student learning for a variety of different visual recognition tasks, because a lightweight student network can be easily trained under the guidance of the high-capacity teacher networks.
- Knowledge distillation can make full use of the different types of knowledge in complex data sources, such as cross-modality data, multi-domain data and multi-task data and low-resolution data, because of flexible teacher–student architectures and knowledge transfer.

7.2 KD in NLP

Conventional language models such as BERT are very time consuming and resource consuming with complex cumbersome structures. Knowledge distillation is extensively studied in the field of natural language processing (NLP), in order to obtain the lightweight, efficient and effective language models. More and more KD methods are proposed for solving the numerous NLP tasks (Liu et al. 2019b; Gordon and Duh 2019; Haidar and Rezagholizadeh 2019; Yang et al. 2020c; Tang et al. 2019; Hu et al. 2018; Sun et al. 2019; Nakashole and Flauger 2017; Jiao et al. 2020; Wang et al. 2018d; Zhou et al. 2019a; Sanh et al. 2019; Turc et al. 2019; Arora et al. 2019; Clark et al. 2019; Kim and Rush 2016; Mou et al. 2016; Liu et al. 2019f; Hahn and Choi 2019; Tan et al. 2019; Kuncoro et al. 2016; Cui et al. 2017; Wei et al. 2019; Freitag et al. 2017; Shakeri et al. 2019; Aguilar et al. 2020; Fu et al. 2021; Zhang et al. 2021b; Chen et al. 2020b; Wang and Du 2021). The existing NLP tasks using KD contain neural machine translation (NMT) (Hahn and Choi 2019; Zhou et al. 2019a; Li et al. 2021; Kim and Rush 2016; Gordon and Duh 2019; Tan et al. 2019; Wei et al. 2019; Freitag et al. 2017; Zhang et al. 2021b), text generation (Chen et al. 2020b; Haidar and Rezagholizadeh 2019), question answering system (Hu et al. 2018; Wang et al. 2018d; Arora et al. 2019; Yang et al. 2020c), event detection (Liu et al. 2019b), document retrieval (Shakeri et al. 2019), text recognition (Wang and Du 2021) and so on. Among these KD-based NLP methods, most of them belong to natural language understanding (NLU), and many of these KD methods for NLU are designed as the task-specific distillation (Tang et al. 2019; Turc et al. 2019; Mou et al. 2016) and multi-task distillation (Liu et al. 2019f; Yang et al. 2020c; Sanh et al. 2019; Clark et al. 2019). In what follows, we describe KD research works for neural machine translation and then for extending a typical multilingual representation model entitled bidirectional encoder

representations from transformers (or BERT) (Devlin et al. 2019) in NLU.

In natural language processing, neural machine translation is the hottest application. However, the existing NMT models with competitive performance is very large. To obtain lightweight NMT, there are many extended knowledge distillation methods for neural machine translation (Hahn and Choi 2019; Zhou et al. 2019a; Kim and Rush 2016; Gordon and Duh 2019; Wei et al. 2019; Freitag et al. 2017; Tan et al. 2019). Recently, Zhou et al. (2019a) empirically proved the better performance of the KD-based non-autoregressive machine translation (NAT) model largely relies on its capacity and the distilled data via knowledge transfer. Gordon and Duh (2019) explained the good performance of sequence-level knowledge distillation from the perspective of data augmentation and regularization. In (Kim and Rush 2016), the effective word-level knowledge distillation is extended to the sequence-level one in the sequence generation scenario of NMT. The sequence generation student model mimics the sequence distribution of the teacher. To overcome the multilingual diversity, Tan et al. (2019) proposed multi-teacher distillation, in which multiple individual models for handling bilingual pairs are teacher and a multilingual model is student. To improve the translation quality, an ensemble of mutiple NMT models as teacher supervise the student model with a data filtering method Freitag et al. (2017). To improve the performance of machine translation and machine reading tasks, (Wei et al. 2019) proposed a novel online knowledge distillation method, which addresses the unstableness of the training process and the decreasing performance on each validation set. In this online KD, the best evaluated model during training is chosen as teacher and updated by any subsequent better model. If the next model had the poor performance, the current teacher model would guide it.

As a multilingual representation model, BERT has attracted attention in natural language understanding (Devlin et al. 2019), but it is also a cumbersome deep model that is not easy to be deployed. To address this problem, several lightweight variations of BERT (called BERT model compression) using knowledge distillation are proposed (Sun et al. 2019; Jiao et al. 2020; Tang et al. 2019; Sanh et al. 2019; Wang et al. 2020a; Liu et al. 2020b; Fu et al. 2021). Sun et al. (2019) proposed patient knowledge distillation for BERT model compression (BERT-PKD), which is used for sentiment classification, paraphrase similarity matching, natural language inference, and machine reading comprehension. In the patient KD method, the feature representations of the [CLS] token from the hint layers of teacher are transferred to the student. To accelerate language inference, Jiao et al. (2020) proposed TinyBERT that is two-stage transformer knowledge distillation. It contains general-domain and task-specific knowledge distillation. For sentence classification and matching, Tang et al. (2019) proposed task-specific

knowledge distillation from the BERT teacher model into a bidirectional long short-term memory network (BiLSTM). In (Sanh et al. 2019), a lightweight student model called Distil-BERT with the same generic structure as BERT is designed and learned on a variety of tasks of NLP. In Aguilar et al. (2020), a simplified student BERT is proposed by using the internal representations of a large teacher BERT via internal distillation.

Furthermore, some typical KD methods for NLP with different perspectives are represented below. For question answering, to improve the efficiency and robustness of machine reading comprehension, Hu et al. (2018) proposed an attention-guided answer distillation method, which fuses generic distillation and answer distillation to avoid confusing answers. For a task-specific distillation (Turc et al. 2019), the performance of knowledge distillation with the interactions among pre-training, distillation and fine-tuning for the compact student model is studied. The proposed pre-trained distillation performs well in sentiment classification, natural language inference, textual entailment. For a multi-task distillation in the context of natural language understanding, Clark et al. (2019) proposed the single-multi born-again distillation, which is based on born-again neural networks (Furlanello et al. 2018). Single-task teachers teach a multi-task student. For multilingual representations, knowledge distillation transfers knowledge among the multilingual word embeddings for bilingual dictionary induction (Nakashole and Flauger 2017). For low-resource languages, knowledge transfer is effective across ensembles of multilingual models (Cui et al. 2017).

Several observations about knowledge distillation for natural language processing are summarized as follows.

- Knowledge distillation provides efficient and effective lightweight language deep models. The large-capacity teacher model can transfer the rich knowledge from a large number of different kinds of language data to train a small student model, so that the student can quickly complete many language tasks with effective performance.
- The teacher–student knowledge transfer can easily and effectively solve many multilingual tasks, considering that knowledge from multilingual models can be transferred and shared by each other.
- In deep language models, the sequence knowledge can be effectively transferred from large networks into small networks.

7.3 KD in Speech Recognition

In the field of speech recognition, deep neural acoustic models have attracted attention and interest due to their powerful performance. However, more and more real-time speech recognition systems are deployed in embedded platforms with limited computational resources and fast response time. The state-of-the-art deep complex models cannot satisfy the requirement of such speech recognition scenarios. To satisfy these requirements, knowledge distillation is widely studied and applied in many speech recognition tasks. There are many knowledge distillation systems for designing lightweight deep acoustic models for speech recognition (Chebotar and Waters 2016; Wong and Gales 2016; Chan et al. 2015; Price et al. 2016; Fukuda et al. 2017; Bai et al. 2019; Ng et al. 2018; Albanie et al. 2018; Lu et al. 2017; Shi et al. 2019a; Roheda et al. 2018; Shi et al. 2019b; Gao et al. 2019; Ghorbani et al. 2018; Takashima et al. 2018; Watanabe et al. 2017; Shi et al. 2019c; Asami et al. 2017; Huang et al. 2018; Shen et al. 2018; Perez et al. 2020; Shen et al. 2019c; Oord et al. 2018; Kwon et al. 2020; Shen et al. 2020). In particular, these KD-based speech recognition applications have spoken language identification (Shen et al. 2018, 2019c, 2020), audio classification (Gao et al. 2019; Perez et al. 2020), text-independent speaker recognition (Ng et al. 2018), speech enhancement (Watanabe et al. 2017), acoustic event detection (Price et al. 2016; Shi et al. 2019a, b), speech synthesis (Oord et al. 2018) and so on.

Most existing knowledge distillation methods for speech recognition, use teacher–student architectures to improve the efficiency and recognition accuracy of acoustic models (Chan et al. 2015; Watanabe et al. 2017; Chebotar and Waters 2016; Shen et al. 2019c; Lu et al. 2017; Shen et al. 2018, 2020; Gao et al. 2019; Shi et al. 2019c, a; Perez et al. 2020). Using a recurrent neural network (RNN) for holding the temporal information from speech sequences, the knowledge from the teacher RNN acoustic model is transferred into a small student DNN model (Chan et al. 2015). Better speech recognition accuracy is obtained by combining multiple acoustic modes. The ensembles of different RNNs with different individual training criteria are designed to train a student model through knowledge transfer (Chebotar and Waters 2016). The learned student model performs well on 2,000-hour large vocabulary continuous speech recognition (LVCSR) tasks in 5 languages. To strengthen the generalization of the spoken language identification (LID) model on short utterances, the knowledge of feature representations of the long utterance-based teacher network is transferred into the short utterance-based student network that can discriminate short utterances and perform well on the short duration utterance-based LID tasks (Shen et al. 2018). To further improve the performance of short utterance-based LID, an interactive teacher–student online distillation learning is proposed to enhance the performance of the feature representations of short utterances (Shen et al. 2019c). The LID performance on short utterances is also improved by distilling internal representation knowledge of teacher on longer utterances into the one of student on short utterances (Shen et al. 2020).

Meanwhile, for audio classification, a multi-level feature distillation method is developed and an adversarial learning strategy is adopted to optimize the knowledge transfer (Gao et al. 2019). To improve noise robust speech recognition, knowledge distillation is employed as the tool of speech enhancement (Watanabe et al. 2017). In Perez et al. (2020), a audio-visual multi-modal knowledge distillation method is proposed. knowledge is transferred from the teacher models on visual and acoustic data into a student model on audio data. In essence, this distillation shares the cross-modal knowledge among the teachers and students (Perez et al. 2020; Albanie et al. 2018; Roheda et al. 2018). For efficient acoustic event detection, a quantized distillation method is proposed by using both knowledge distillation and quantization (Shi et al. 2019a). The quantized distillation transfers knowledge from a large CNN teacher model with better detection accuracy into a quantized RNN student model.

Unlike most existing traditional frame-level KD methods, sequence-level KD can perform better in some sequence models for speech recognition, such as connectionist temporal classification (CTC) (Wong and Gales 2016; Takashima et al. 2018; Huang et al. 2018). In (Huang et al. 2018), sequence-level KD is introduced into connectionist temporal classification, in order to match an output label sequence used in the training of teacher model and the input speech frames used in distillation. In Wong and Gales (2016), the effect of speech recognition performance on frame-level and sequence-level student-teacher training is studied and a new sequence-level student-teacher training method is proposed. The teacher ensemble is constructed by using sequence-level combination instead of frame-level combination. To improve the performance of unidirectional RNN-based CTC for real-time speech recognition, the knowledge of a bidirectional LSTM-based CTC teacher model is transferred into a unidirectional LSTM-based CTC student model via frame-level KD and sequence-level KD (Takashima et al. 2018).

Moreover, knowledge distillation can be used to solve some special issues in speech recognition (Bai et al. 2019; Asami et al. 2017; Ghorbani et al. 2018). To overcome overfitting issue of DNN acoustic models when data are scarce, knowledge distillation is employed as a regularization way to train adapted model with the supervision of the source model (Asami et al. 2017). The final adapted model achieves better performance on three real acoustic domains. To overcome the degradation of the performance of non-native speech recognition, an advanced multi-accent student model is trained by distilling knowledge from the multiple accent-specific RNN-CTC models (Ghorbani et al. 2018). In essence, knowledge distillation in (Asami et al. 2017; Ghorbani et al. 2018) realizes the cross-domain knowledge transfer. To solve the complexity of fusing the external language model (LM) into sequence-to-sequence model (Seq2seq) for speech recognition, knowledge distillation is employed as an effective tool to integrate a LM (teacher) into Seq2seq model (student) (Bai et al. 2019). The trained Seq2seq model can reduce character error rates in sequence-to-sequence speech recognition.

In summary, several observations on knowledge distillation-based speech recognition can be concluded as follows.

- The lightweight student model can satisfy the practical requirements of speech recognition, such as real-time responses, use of limited resources and high recognition accuracy.
- Many teacher–student architectures are built on RNN models because of the temporal property of speech sequences. In general, the RNN models are chosen as the teacher, which can well preserve and transfer the temporal knowledge from real acoustic data to a student model.
- Sequence-level knowledge distillation can be well applied to sequence models with good performance. In fact, the frame-level KD always uses the response-based knowledge, but sequence-level KD usually transfers the feature-based knowledge from hint layers of teacher models.
- Knowledge distillation using teacher–student knowledge transfer can easily solve the cross-domain or cross-modal speech recognition in applications such as multi-accent and multilingual speech recognition.

7.4 KD in Other Applications

The full and correct leverages of external knowledge, such as in a user review or in images, play a very important role in the effectiveness of deep recommendation models. Reducing the complexity and improving the efficiency of deep recommendation models is also very necessary. Recently, knowledge distillation has been successfully applied in recommender systems for deep model compression and acceleration (Chen et al. 2018b; Tang and Wang 2018; Pan et al. 2019). In (Tang and Wang 2018), knowledge distillation is first introduced into the recommender systems and called ranking distillation because the recommendation is expressed as a ranking problem. Chen et al. (2018b) proposed an adversarial knowledge distillation method for efficient recommendation. A teacher as the right review predication network supervises the student as user-item prediction network (generator). The student learning is adjusted by adversarial adaption between teacher and student networks. Unlike distillation in Chen et al. (2018b), Tang and Wang (2018), Pan et al. (2019) designed a enhanced collaborative denoising autoencoder (ECAE) model for recommender systems via knowledge distillation to capture useful knowledge from user feedbacks and to reduce noise. The unified ECAE framework contains a generation network, a retraining network and a distillation layer that transfers knowledge and reduces noise from the generation network.

Using the natural characteristic of knowledge distillation with teacher–student architectures, knowledge distillation is used as an effective strategy to solve adversarial attacks or perturbations of deep models (Papernot et al. 2016; Ross and Doshi-Velez 2018; Goldblum et al. 2020; Gil et al. 2019) and the issue of the unavailable data due to the privacy, confidentiality and security concerns (Lopes et al. 2017; Papernot et al. 2017; Wang et al. 2019a; Bai et al. 2020; Vongkulbhisal et al. 2019). To be specific, the perturbations of the adversarial samples can be overcome by the robust outputs of the teacher networks via distillation (Ross and Doshi-Velez 2018; Papernot et al. 2016). To avoid exposing the private data, multiple teachers access subsets of the sensitive or unlabelled data and supervise the student (Papernot et al. 2017; Vongkulbhisal et al. 2019). To address the issue of privacy and security, the data to train the student network is generated by using the layer activations or layer spectral activations of the teacher network via data-free distillation (Lopes et al. 2017). To protect data privacy and prevent intellectual piracy, Wang et al. (2019a) proposed a private model compression framework via knowledge distillation. The student model is applied to public data while the teacher model is applied to both sensitive and public data. This private knowledge distillation adopts privacy loss and batch loss to further improve privacy. To consider the compromise between privacy and performance, Bai et al. (2020) developed a few shot network compression method via a novel layer-wise knowledge distillation with few samples per class. Of course, there are other special interesting applications of knowledge distillation, such as neural architecture search (Macko et al. 2019; Bashivan et al. 2019), interpretability of deep neural networks (Liu et al. 2018b), and federated learning (Bistritz et al. 2020; Lin et al. 2020; Seo et al. 2020; He et al. 2020a).

8 Conclusion and Discussion

Knowledge distillation and its applications have aroused considerable attention in recent few years. In this paper, we present a comprehensive review on knowledge distillation, from the perspectives of knowledge, distillation schemes, teacher–student architectures, distillation algorithms, performance comparison and applications. Below, we discuss the challenges of knowledge distillation and provide some insights on the future research of knowledge distillation.

8.1 Challenges

For knowledge distillation, the key is to 1) extract rich knowledge from the teacher and 2) to transfer the knowledge from the teacher to guide the training of the student. Therefore, we discuss the challenges in knowledge distillation from the followings aspects: the quality of knowledge, the types of distillation, the design of the teacher–student architectures, and the theory behind knowledge distillation.

Most KD methods leverage a combination of different kinds of knowledge, including response-based, feature-based, and relation-based knowledge. Therefore, it is important to know the influence of each individual type of knowledge and to know how different kinds of knowledge help each other in a complementary manner. For example, the response-based knowledge has a similar motivation to label smoothing and the model regularization (Kim and Kim 2017; Muller et al. 2019; Ding et al. 2019); The featured-based knowledge is often used to mimic the intermediate process of the teacher and the relation-based knowledge is used to capture the relationships across different samples. To this end, it is still challenging to model different types of knowledge in a unified and complementary framework. For example, the knowledge from different hint layers may have different influences on the training of the student model: 1) response-based knowledge is from the last layer; 2) feature-based knowledge from the deeper hint/guided layers may suffer from over-regularization (Romero et al. 2015).

How to transfer the rich knowledge from the teacher to a student is a key step in knowledge distillation. Generally, the existing distillation methods can be categorized into offline distillation, online distillation and self distillation. Offline distillation is usually used to transfer knowledge from a complex teacher model, while the teacher model and the student model are comparable in the settings of online distillation and self distillation. To improve the efficacy of knowledge transfer, the relationships between the model complexity and existing distillation schemes or other novel distillation schemes (Sun et al. 2021) should be further investigated.

Currently, most KD methods focus on new types of knowledge or distillation loss functions, leaving the design of the teacher–student architectures poorly investigated (Nowak and Corso 2018; Crowley et al. 2018; Kang et al. 2020; Liu et al. 2019i; Ashok et al. 2018; Liu et al. 2019a). In fact, apart from the knowledge and distillation algorithms, the relationship between the structures of the teacher and the student also significantly influences the performance of knowledge distillation. For example, on one hand, some recent works find that the student model can learn little from some teacher models due to the model capacity gap between the teacher model and the student model (Zhang et al. 2019b; Kang et al. 2020); On the other hand, from some early theoretical analysis on the capacity of neural networks, shallow networks are capable of learning the same representation as deep neural networks (Ba and Caruana 2014). Therefore, the design of an effective student model or construction of a proper teacher model are still challenging problems in knowledge distillation.

Despite a huge number of the knowledge distillation methods and applications, the understanding of knowledge distillation including theoretical explanations and empiri-

cal evaluations remains insufficient (Lopez-Paz et al. 2016; Phuong and Lampert 2019a; Cho and Hariharan 2019). For example, distillation can be viewed as a form of learning with privileged information (Lopez-Paz et al. 2016). The assumption of linear teacher and student models enables the study of the theoretical explanations of characteristics of the student learning via distillation (Phuong and Lampert 2019a). Furthermore, some empirical evaluations and analysis on the efficacy of knowledge distillation were performed by Cho and Hariharan (2019). However, a deep understanding of generalizability of knowledge distillation, especially how to measure the quality of knowledge or the quality of the teacher–student architecture, is still very difficult to attain.

8.2 Future Directions

In order to improve the performance of knowledge distillation, the most important factors include what kind of teacher–student network architecture, what kind of knowledge is learned from the teacher network, and where is distilled into the student network.

The model compression and acceleration methods for deep neural networks usually fall into four different categories, namely parameter pruning and sharing, low-rank factorization, transferred compact convolutional filters and knowledge distillation (Cheng et al. 2018). In existing knowledge distillation methods, there are only a few related works discussing the combination of knowledge distillation and other kinds of compressing methods. For example, quantized knowledge distillation, which can be seen as a parameter pruning method, integrates network quantization into the teacher–student architectures (Polino et al. 2018; Mishra and Marr 2018; Wei et al. 2018). Therefore, to learn efficient and effective lightweight deep models for the deployment on portable platforms, the hybrid compression methods via both knowledge distillation and other compressing techniques are necessary, since most compressing techniques require a re-training/fine-tuning process. Furthermore, how to decide the proper orders for applying different compressing methods will be an interesting topic for future study.

Apart from model compression for acceleration for deep neural networks, knowledge distillation also can be used in other problems because of the natural characteristics of knowledge transfer on the teacher–student architecture. Recently, knowledge distillation has been applied to the data privacy and security (Wang et al. 2019a), adversarial attacks of deep models (Papernot et al. 2016), cross-modalities (Gupta et al. 2016), multiple domains (Asami et al. 2017), catastrophic forgetting (Lee et al. 2019b), accelerating learning of deep models (Chen et al. 2016), efficiency of neural architecture search (Bashivan et al. 2019), self-supervision (Noroozi et al. 2018), and data augmentation (Lee et al. 2019a; Gordon and Duh 2019). Another interesting exam-

ple is that the knowledge transfer from the small teacher networks to a large student network can accelerate the student learning (Chen et al. 2016). This is very quite different from vanilla knowledge distillation. The feature representations learned from unlabelled data by a large model can also supervise the target model via distillation (Noroozi et al. 2018). To this end, the extensions of knowledge distillation for other purposes and applications might be a meaningful future direction.

The learning of knowledge distillation is similar to the human beings learning. It can be practicable to popularize the knowledge transfer to the classic and traditional machine learning methods (Zhou et al. 2019b; Gong et al. 2018; You et al. 2018; Gong et al. 2017). For example, traditional two-stage classification is felicitous cast to a single teacher single student problem based on the idea of knowledge distillation (Zhou et al. 2019b). Furthermore, knowledge distillation can be flexibly deployed to various excellent learning schemes, such as the adversarial learning (Liu et al. 2018), auto machine learning (Macko et al. 2019; Fakoor et al. 2020), label noise filtering learning (Xia et al. 2018), lifelong learning (Zhai et al. 2019), and reinforcement learning (Ashok et al. 2018; Xu et al. 2020c; Zhao and Hospedales 2020). Therefore, it will be useful to integrate knowledge distillation with other learning schemes for practical challenges in the future.

References

Aditya, S., Saha, R., Yang, Y., & Baral, C. (2019). Spatial knowledge distillation to aid visual reasoning. In *WACV*.

Aguilar, G., Ling, Y., Zhang, Y., Yao, B., Fan, X., & Guo, E. (2020). Knowledge distillation from internal representations. In *AAAI*.

Aguinaldo, A., Chiang, P. Y., Gain, A., Patil, A., Pearson, K., & Feizi, S. (2019). Compressing gans using knowledge distillation. arXiv preprint arXiv:1902.00159.

Ahn, S., Hu, S., Damianou, A., Lawrence, N. D., & Dai, Z. (2019). Variational information distillation for knowledge transfer. In *CVPR*.

Albanie, S., Nagrani, A., Vedaldi, A., & Zisserman, A. (2018). Emotion recognition in speech using cross-modal transfer in the wild. In *ACM MM*.

Allen-Zhu, Z., Li, Y., & Liang, Y. (2019). Learning and generalization in overparameterized neural networks, going beyond two layers. In *NeurIPS*.

Anil, R., Pereyra, G., Passos, A., Ormandi, R., Dahl, G. E., & Hinton, G. E. (2018). Large scale distributed neural network training through online distillation. In *ICLR*.

Arora, S., Cohen, N., & Hazan, E. (2018). On the optimization of deep networks: Implicit acceleration by overparameterization. In *ICML*.

Arora, S., Khapra, M. M., & Ramaswamy, H. G. (2019). On knowledge distillation from complex networks for response prediction. In *NAACL-HLT*.

Asami, T., Masumura, R., Yamaguchi, Y., Masataki, H., & Aono, Y. (2017). Domain adaptation of dnn acoustic models using knowledge distillation. In *ICASSP*.

Ashok, A., Rhinehart, N., Beainy, F., & Kitani, K. M. (2018). N2N learning: Network to network compression via policy gradient reinforcement learning. In *ICLR*.

Asif, U., Tang, J. & Harrer, S. (2020). Ensemble knowledge distillation for learning improved and efficient networks. In *ECAI*.

Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep? In *NeurIPS*.

Bagherinezhad, H., Horton, M., Rastegari, M., & Farhadi, A. (2018). Label refinery: Improving imagenet classification through label progression. arXiv preprint arXiv:1805.02641.

Bai, H., Wu, J., King, I., & Lyu, M. (2020). Few shot network compression via cross distillation. In *AAAI*.

Bai, Y., Yi, J., Tao, J., Tian, Z., & Wen, Z. (2019). Learn spelling from teachers: transferring knowledge from language models to sequence-to-sequence speech recognition. In *Interspeech*.

Bashivan, P., Tensen, M., & DiCarlo, J. J. (2019). Teacher guided architecture search. In *ICCV*.

Belagiannis, V., Farshad, A., & Galasso, F. (2018). Adversarial network compression. In *ECCV*.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE TPAMI, 35*(8), 1798–1828.

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2020). Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*.

Bhardwaj, S., Srinivasan, M., & Khapra, M. M. (2019). Efficient video classification using fewer frames. In *CVPR*.

Bistritz, I., Mann, A., & Bambos, N. (2020). Distributed Distillation for On-Device Learning. In *NeurIPS*.

Bohdal, O., Yang, Y., & Hospedales, T. (2020). Flexible Dataset Distillation: Learn Labels Instead of Images. arXiv preprint arXiv:2006.08572.

Boo, Y., Shin, S., Choi, J., & Sung, W. (2021). Stochastic precision ensemble: self-knowledge distillation for quantized deep neural networks. In *AAAI*.

Brutzkus, A., & Globerson, A. (2019). Why do Larger Models Generalize Better? A Theoretical Perspective via the XOR Problem. In *ICML*.

Bucilua, C., Caruana, R. & Niculescu-Mizil, A. (2006). Model compression. In *SIGKDD*.

Caccia, M., Rodriguez, P., Ostapenko, O., Normandin, F., Lin, M., Caccia, L., Laradji, I., Rish, I., Lacoste, A., Vazquez D., & Charlin, L. (2020). Online Fast Adaptation and Knowledge Accumulation (OSAKA): a New Approach to Continual Learning. In *NeurIPS*.

Chan, W., Ke, N. R., & Lane, I. (2015). Transferring knowledge from a RNN to a DNN. arXiv preprint arXiv:1504.01483.

Chawla, A., Yin, H., Molchanov, P., & Alvarez, J. (2021). Data-Free Knowledge Distillation for Object Detection. In *WACV*.

Chebotar, Y. & Waters, A. (2016). Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech*.

Chen, D., Mei, J. P., Wang, C., Feng, Y. & Chen, C. (2020a). Online knowledge distillation with diverse peers. In *AAAI*.

Chen, D., Mei, J. P., Zhang, Y., Wang, C., Wang, Z., Feng, Y., & Chen, C. (2021). Cross-layer distillation with semantic calibration. In *AAAI*.

Chen, G., Choi, W., Yu, X., Han, T., & Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. In *NeurIPS*.

Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., &Tian, Q. (2019a). Data-free learning of student networks. In *ICCV*.

Chen, H., Wang, Y., Xu, C., Xu, C., & Tao, D. (2021). Learning student networks via feature embedding. *IEEE TNNLS, 32*(1), 25–35.

Chen, T., Goodfellow, I. & Shlens, J. (2016). Net2net: Accelerating learning via knowledge transfer. In *ICLR*.

Chen, W. C., Chang, C. C. & Lee, C. R. (2018a). Knowledge distillation with feature maps for image classification. In *ACCV*.

Chen, X., Zhang, Y., Xu, H., Qin, Z., & Zha, H. (2018b). Adversarial distillation for efficient recommendation with external knowledge. *ACM TOIS, 37*(1), 1–28.

Chen, X., Su, J., & Zhang, J. (2019b). A two-teacher tramework for knowledge distillation. In *ISNN*.

Chen, Y., Wang, N., & Zhang, Z. (2018c). Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In *AAAI*.

Chen, Y. C., Gan, Z., Cheng, Y., Liu, J., & Liu, J. (2020b). Distilling knowledge learned in BERT for text generation. In *ACL*.

Chen, Y.C., Lin, Y. Y., Yang, M. H., Huang, J. B. (2019c). Crdoco: Pixel-level domain transfer with cross-domain consistency. In *CVPR*.

Chen, Z., & Liu, B. (2018). Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 12*(3), 1–207.

Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., & Heng, P. A. (2020c). A multi-task mean teacher for semi-supervised shadow detection. In *CVPR*.

Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2018). Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine, 35*(1), 126–136.

Cheng, X., Rao, Z., Chen, Y., & Zhang, Q. (2020). Explaining knowledge distillation by quantifying the knowledge. In *CVPR*.

Cho, J. H. & Hariharan, B. (2019). On the efficacy of knowledge distillation. In *ICCV*.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *CVPR*.

Chung, I., Park, S., Kim, J. & Kwak, N. (2020). Feature-map-level online adversarial knowledge distillation. In *ICML*.

Clark, K., Luong, M. T., Khandelwal, U., Manning, C. D. & Le, Q. V. (2019). Bam! born-again multi-task networks for natural language understanding. In *ACL*.

Courbariaux, M., Bengio, Y. & David, J. P. (2015). Binaryconnect: Training deep neural networks with binary weights during propagations. In *NeurIPS*.

Crowley, E. J., Gray, G. & Storkey, A. J. (2018). Moonshine: Distilling with cheap convolutions. In *NeurIPS*.

Cui, J., Kingsbury, B., Ramabhadran, B., Saon, G., Sercu, T., Audhkhasi, K., et al. (2017). Knowledge distillation across ensembles of multilingual models for low-resource languages. In *ICASSP*.

Cui, Z., Song, T., Wang, Y., & Ji, Q. (2020). Knowledge augmented deep neural networks for joint facial expression and action unit recognition. In *NeurIPS*.

Cun, X., & Pun, C. M. (2020). Defocus blur detection via depth distillation. In *ECCV*.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*.

Denton, E. L., Zaremba, W., Bruna, J., LeCun, Y. & Fergus, R. (2014). Exploiting linear structure within convolutional networks for efficient evaluation. In *NeurIPS*.

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT* .

Ding, Q., Wu, S., Sun, H., Guo, J. & Xia, S. T. (2019). Adaptive regularization of labels. arXiv preprint arXiv:1908.05474.

Do, T., Do, T. T., Tran, H., Tjiputra, E. & Tran, Q. D. (2019). Compact trilinear interaction for visual question answering. In *ICCV*.

Dong, X. & Yang, Y. (2019). Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *ICCV*.

Dou, Q., Liu, Q., Heng, P. A., & Glocker, B. (2020). Unpaired multimodal segmentation via knowledge distillation. *IEEE TMI, 39*(7), 2415–2425.

Du, S., You, S., Li, X., Wu, J., Wang, F., Qian, C., & Zhang, C. (2020). Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. In *NeurIPS*.

Duong, C. N., Luu, K., Quach, K. G. & Le, N. (2019.) ShrinkTeaNet: Million-scale lightweight face recognition via shrinking teacher-student networks. arXiv preprint arXiv:1905.10620.

Fakoor, R., Mueller, J. W., Erickson, N., Chaudhari, P., & Smola, A. J. (2020). Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation. In *NeurIPS*.

Flennerhag, S., Moreno, P. G., Lawrence, N. D. & Damianou, A. (2019). Transferring knowledge across learning processes. In *ICLR*.

Freitag, M., Al-Onaizan, Y. & Sankaran, B. (2017). Ensemble distillation for neural machine translation. arXiv preprint arXiv:1702.01802.

Fu, H., Zhou, S., Yang, Q., Tang, J., Liu, G., Liu, K., & Li, X. (2021). LRC-BERT: Latent-representation Contrastive Knowledge Distillation for Natural Language Understanding. In *AAAI*.

Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J. & Ramabhadran, B. (2017). Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*.

Furlanello, T., Lipton, Z., Tschannen, M., Itti, L. & Anandkumar, A. (2018). Born again neural networks. In *ICML*.

Gao, L., Mi, H., Zhu, B., Feng, D., Li, Y., & Peng, Y. (2019). An adversarial feature distillation method for audio classification. *IEEE Access, 7,* 105319–105330.

Gao, M., Wang, Y., & Wan, L. (2021). Residual error based knowledge distillation. *Neurocomputing, 433,* 154–161.

Gao, Z., Chung, J., Abdelrazek, M., Leung, S., Hau, W. K., Xian, Z., et al. (2020). Privileged modality distillation for vessel border detection in intracoronary imaging. *IEEE TMI, 39*(5), 1524–1534.

Garcia, N. C., Morerio, P. & Murino, V. (2018). Modality distillation with multiple stream networks for action recognition. In *ECCV*.

Ge, S., Zhao, S., Li, C., & Li, J. (2018). Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE TIP, 28*(4), 2051–2062.

Ge, S., Zhao, S., Li, C., Zhang, Y., & Li, J. (2020). Efficient low-resolution face recognition via bridge distillation. *IEEE TIP, 29,* 6898–6908.

Ghorbani, S., Bulut, A. E. & Hansen, J. H. (2018). Advancing multi-accented lstm-ctc speech recognition using a domain specific student-teacher learning paradigm. In *SLTW*.

Gil, Y., Chai, Y., Gorodissky, O. & Berant, J. (2019). White-to-black: Efficient distillation of black-box adversarial attacks. In *NAACL-HLT*.

Goldblum, M., Fowl, L., Feizi, S. & Goldstein, T. (2020). Adversarially robust distillation. In *AAAI*.

Gong, C., Chang, X., Fang, M. & Yang, J. (2018). Teaching semi-supervised classifier via generalized distillation. In *IJCAI*.

Gong, C., Tao, D., Liu, W., Liu, L., & Yang, J. (2017). Label propagation via teaching-to-learn and learning-to-teach. *TNNLS, 28*(6), 1452–1465.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *NeurIPS*.

Gordon, M. A. & Duh, K. (2019). Explaining sequence-level knowledge distillation as data-augmentation for neural machine translation. arXiv preprint arXiv:1912.03334.

Gu, J., & Tresp, V. (2020). Search for better students to learn distilled knowledge. In *ECAI*.

Guan, Y., Zhao, P., Wang, B., Zhang, Y., Yao, C., Bian, K., & Tang, J. (2020). Differentiable feature aggregation search for knowledge distillation. In *ECCV*.

Guo, Q., Wang, X., Wu, Y., Yu, Z., Liang, D., Hu, X., & Luo, P. (2020). Online knowledge distillation via collaborative learning. In *CVPR*.

Gupta, S., Hoffman, J. & Malik, J. (2016). Cross modal distillation for supervision transfer. In *CVPR*.

Hahn, S. & Choi, H. (2019). Self-knowledge distillation in natural language processing. In *RANLP*.

Haidar, M. A. & Rezagholizadeh, M. (2019). Textkd-gan: Text generation using knowledge distillation and generative adversarial networks. In *Canadian conference on artificial intelligence*.

Han, S., Pool, J., Tran, J. & Dally, W. (2015). Learning both weights and connections for efficient neural network. In *NeurIPS*.

Hao, W., & Zhang, Z. (2019). Spatiotemporal distilled dense-connectivity network for video action recognition. *Pattern Recognition, 92,* 13–24.

Haroush, M., Hubara, I., Hoffer, E., & Soudry, D. (2020). The knowledge within: Methods for data-free model compression. In *CVPR*.

He, C., Annavaram, M., & Avestimehr, S. (2020a). Group knowledge transfer: Federated learning of large CNNs at the edge. In *NeurIPS*.

He, F., Liu, T., & Tao, D. (2020b). Why resnet works? residuals generalize. *IEEE TNNLS, 31*(12), 5349–5362.

He, K., Zhang, X., Ren, S. & Sun, J. (2016) Deep residual learning for image recognition. In *CVPR*.

He, T., Shen, C., Tian, Z., Gong, D., Sun, C. & Yan, Y. (2019). Knowledge adaptation for efficient semantic segmentation. In *CVPR*.

Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., & Choi, J. Y. (2019a). A comprehensive overhaul of feature distillation. In *ICCV*.

Heo, B., Lee, M., Yun, S. & Choi, J. Y. (2019b). Knowledge distillation with adversarial samples supporting decision boundary. In *AAAI*.

Heo, B., Lee, M., Yun, S. & Choi, J. Y. (2019c). Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *AAAI*.

Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hoffman, J., Gupta, S. & Darrell, T. (2016). Learning with side information through modality hallucination. In *CVPR*.

Hong, W. & Yu, J. (2019). Gan-knowledge distillation for one-stage object detection. arXiv preprint arXiv:1906.08467.

Hou, Y., Ma, Z., Liu, C. & Loy, CC. (2019). Learning lightweight lane detection cnns by self attention distillation. In *ICCV*.

Hou, Y., Ma, Z., Liu, C., Hui, T. W., & Loy, C. C. (2020). Inter-Region Affinity Distillation for Road Marking Segmentation. In *CVPR*.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hu, H., Xie, L., Hong, R., & Tian, Q. (2020). Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In *CVPR*.

Hu, M., Peng, Y., Wei, F., Huang, Z., Li, D., Yang, N., et al. (2018). Attention-guided answer distillation for machine reading comprehension. In *EMNLP*.

Huang, G., Liu, Z., Van, Der Maaten, L. & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *CVPR*.

Huang, M., You, Y., Chen, Z., Qian, Y. & Yu, K. (2018). Knowledge distillation for sequence model. In *Interspeech*.

Huang, Z. & Wang, N. (2017). Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219.

Huang, Z., Zou, Y., Bhagavatula, V., & Huang, D. (2020). Comprehensive attention self-distillation for weakly-supervised object detection. In *NeurIPS*.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*

Jang, Y., Lee, H., Hwang, S. J. & Shin, J. (2019). Learning what and where to transfer. In *ICML*.

Ji, G., & Zhu, Z. (2020). Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. In *NeurIPS*.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., et al. (2020). Tinybert: Distilling bert for natural language understanding. In *EMNLP*.

Jin, X., Peng, B., Wu, Y., Liu, Y., Liu, J., Liang, D., Yan, J., & Hu, X. (2019). Knowledge distillation via route constrained optimization. In *ICCV*.

Kang, M., Mun, J. & Han, B. (2020). Towards oracle knowledge distillation with neural architecture search. In *AAAI*.

Kim, J., Park, S. & Kwak, N. (2018). Paraphrasing complex network: Network compression via factor transfer. In *NeurIPS*.

Kim, J., Bhalgat, Y., Lee, J., Patel, C., & Kwak, N. (2019a). QKD: Quantization-aware Knowledge Distillation. arXiv preprint arXiv:1911.12491.

Kim, J., Hyun, M., Chung, I. & Kwak, N. (2019b). Feature fusion for online mutual knowledge distillation. In *ICPR*.

Kim, S. W. & Kim, H. E. (2017). Transferring knowledge to smaller network with class-distance loss. In *ICLRW*.

Kim, Y., Rush & A. M. (2016). Sequence-level knowledge distillation. In *EMNLP*.

Kimura, A., Ghahramani, Z., Takeuchi, K., Iwata, T. & Ueda, N. (2018). Few-shot learning of neural networks from scratch by pseudo example optimization. In *BMVC*.

Kwon, K., Na, H., Lee, H., & Kim, N. S. (2020). Adaptive knowledge distillation based on entropy. In *ICASSP*.

Kong, H., Zhao, J., Tu, X., Xing, J., Shen, S. & Feng, J. (2019). Cross-resolution face recognition via prior-aided face hallucination and residual knowledge distillation. arXiv preprint arXiv:1905.10777.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NeurIPS*.

Kuncoro, A., Ballesteros, M., Kong, L., Dyer, C. & Smith, N. A. (2016). Distilling an ensemble of greedy dependency parsers into one mst parser. In *EMNLP*.

Kundu, J. N., Lakkakula, N. & Babu, R. V. (2019). Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *CVPR*.

Lai, K. H., Zha, D., Li, Y., & Hu, X. (2020). Dual policy distillation. In *IJCAI*.

Lan, X., Zhu, X., & Gong, S. (2018). Self-referenced deep learning. In *ACCV*.

Lee, H., Hwang, S. J. & Shin, J. (2019a). Rethinking data augmentation: Self-supervision and self-distillation. arXiv preprint arXiv:1910.05872.

Lee, K., Lee, K., Shin, J. & Lee, H. (2019b). Overcoming catastrophic forgetting with unlabeled data in the wild. In *ICCV*.

Lee, K., Nguyen, L. T. & Shim, B. (2019c). Stochasticity and skip connections improve knowledge transfer. In *AAAI*.

Lee, S. & Song, B. (2019). Graph-based knowledge distillation by multi-head attention network. In *BMVC*.

Lee, S. H., Kim, D. H. & Song, B. C. (2018). Self-supervised knowledge distillation using singular value decomposition. In *ECCV*.

Li, B., Wang, Z., Liu, H., Du, Q., Xiao, T., Zhang, C., & Zhu, J. (2021). Learning light-weight translation models from deep transformer. In *AAAI*.

Li, C., Peng, J., Yuan, L., Wang, G., Liang, X., Lin, L., & Chang, X. (2020a). Blockwisely supervised neural architecture search with knowledge distillation. In *CVPR*.

Li, G., Zhang, J., Wang, Y., Liu, C., Tan, M., Lin, Y., Zhang, W., Feng, J., & Zhang, T. (2020b). Residual distillation: Towards portable deep neural networks without shortcuts. In *NeurIPS*.

Li, J., Fu, K., Zhao, S., & Ge, S. (2019). Spatiotemporal knowledge distillation for efficient estimation of aerial video saliency. *IEEE TIP, 29,* 1902–1914.

Li, M., Lin, J., Ding, Y., Liu, Z., Zhu, J. Y., & Han, S. (2020c). Gan compression: Efficient architectures for interactive conditional gans. In *CVPR*.

Li, Q., Jin, S. & Yan, J. (2017). Mimicking very efficient network for object detection. In *CVPR*.

Li, T., Li, J., Liu, Z., & Zhang, C. (2020d). Few sample knowledge distillation for efficient network compression. In *CVPR*.

Li, X., Wu, J., Fang, H., Liao, Y., Wang, F., & Qian, C. (2020e). Local correlation consistency for knowledge distillation. In *ECCV*.

Li, Z., & Hoiem, D. (2017). Learning without forgetting. *IEEE TPAMI, 40*(12), 2935–2947.

Lin, T., Kong, L., Stich, S. U., & Jaggi, M. (2020). Ensemble distillation for robust model fusion in federated learning. In *NeurIPS*.

Liu, I. J., Peng, J. & Schwing, A. G. (2019a). Knowledge flow: Improve upon your teachers. In *ICLR*.

Liu, J., Chen, Y. & Liu, K. (2019b). Exploiting the ground-truth: An adversarial imitation based knowledge distillation approach for event detection. In *AAAI*.

Liu, J., Wen, D., Gao, H., Tao, W., Chen, T. W., Osa, K., et al. (2019c). Knowledge representing: efficient, sparse representation of prior knowledge for knowledge distillation. In *CVPRW*.

Liu, P., King, I., Lyu, M. R., & Xu, J. (2019d). DDFlow: Learning optical flow with unlabeled data distillation. In *AAAI*.

Liu, P., Liu, W., Ma, H., Mei, T. & Seok, M. (2020a). Ktan: knowledge transfer adversarial network. In *IJCNN*.

Liu, Q., Xie, L., Wang, H., Yuille & A. L. (2019e). Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *ICCV*.

Liu, R., Fusi, N. & Mackey, L. (2018). Model compression with generative adversarial networks. arXiv preprint arXiv:1812.02271.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Deng, H., & Ju, Q. (2020b). FastBERT: a self-distilling BERT with adaptive inference time. In *ACL*.

Liu, X., Wang, X. & Matwin, S. (2018b). Improving the interpretability of deep neural networks with knowledge distillation. In *ICDMW*.

Liu, X., He, P., Chen, W. & Gao, J. (2019f). Improving multi-task deep neural networks via knowledge distillation for natural language understanding. arXiv preprint arXiv:1904.09482.

Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y. & Duan, Y. (2019g). Knowledge distillation via instance relationship graph. In *CVPR*.

Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z. & Wang, J. (2019h). Structured knowledge distillation for semantic segmentation. In *CVPR*.

Liu, Y., Jia, X., Tan, M., Vemulapalli, R., Zhu, Y., Green, B., et al. (2019i). Search to distill: Pearls are everywhere but not the eyes. In *CVPR*.

Liu, Y., Zhang, W., & Wang, J. (2020c). Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing, 415,* 106–113.

Lopes, R. G., Fenu, S. & Starner, T. (2017). Data-free knowledge distillation for deep neural networks. In *NeurIPS*.

Lopez-Paz, D., Bottou, L., Schölkopf, B. & Vapnik, V. (2016). Unifying distillation and privileged information. In *ICLR*.

Lu, L., Guo, M. & Renals, S. (2017). Knowledge distillation for small-footprint highway networks. In *ICASSP*.

Luo, P., Zhu, Z., Liu, Z., Wang, X. & Tang, X. (2016). Face model compression by distilling knowledge from neurons. In *AAAI*.

Luo, S., Pan, W., Wang, X., Wang, D., Tang, H., & Song, M. (2020). Collaboration by competition: Self-coordinated knowledge amalgamation for multi-talent student learning. In *ECCV*.

Luo, S., Wang, X., Fang, G., Hu, Y., Tao, D., & Song, M. (2019). Knowledge amalgamation from heterogeneous networks by common feature learning. In *IJCAI*.

Luo, Z., Hsieh, J. T., Jiang, L., Carlos Niebles, J.& Fei-Fei, L. (2018). Graph distillation for action detection with privileged modalities. In *ECCV*.

Macko, V., Weill, C., Mazzawi, H. & Gonzalvo, J. (2019). Improving neural architecture search image classifiers via ensemble learning. In *NeurIPS workshop*.

Ma, J., & Mei, Q. (2019). Graph representation learning via multi-task knowledge distillation. arXiv preprint arXiv:1911.05700.

Ma, N., Zhang, X., Zheng, H. T., & Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient CNN architecture design. In *ECCV*.

Meng, Z., Li, J., Zhao, Y. & Gong, Y. (2019). Conditional teacher-student learning. In *ICASSP*.

Micaelli, P. & Storkey, A. J. (2019). Zero-shot knowledge transfer via adversarial belief matching. In *NeurIPS*.

Minami, S., Hirakawa, T., Yamashita, T. & Fujiyoshi, H. (2019). Knowledge transfer graph for deep collaborative learning. arXiv preprint arXiv:1909.04286.

Mirzadeh, S. I., Farajtabar, M., Li, A. & Ghasemzadeh, H. (2020). Improved knowledge distillation via teacher assistant. In *AAAI*.

Mishra, A. & Marr, D. (2018). Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In *ICLR*.

Mobahi, H., Farajtabar, M., & Bartlett, P. L. (2020). Self-distillation amplifies regularization in hilbert space. In *NeurIPS*.

Mou, L., Jia, R., Xu, Y., Li, G., Zhang, L. & Jin, Z. (2016). Distilling word embeddings: An encoding approach. In *CIKM*.

Mukherjee, P., Das, A., Bhunia, A. K. & Roy, P. P. (2019). Cogni-net: Cognitive feature learning through deep visual perception. In *ICIP*.

Mullapudi, R. T., Chen, S., Zhang, K., Ramanan, D. & Fatahalian, K. (2019). Online model distillation for efficient video inference. In *ICCV*.

Muller, R., Kornblith, S. & Hinton, G. E. (2019). When does label smoothing help? In *NeurIPS*.

Mun, J., Lee, K., Shin, J. & Han, B. (2018). Learning to specialize with knowledge distillation for visual question answering. In *NeurIPS*.

Munjal, B., Galasso, F. & Amin, S. (2019). Knowledge distillation for end-to-end person search. In *BMVC*.

Nakashole, N. & Flauger, R. (2017). Knowledge distillation for bilingual dictionary induction. In *EMNLP*.

Nayak, G. K., Mopuri, K. R., & Chakraborty, A. (2021). Effectiveness of arbitrary transfer sets for data-free knowledge distillation. In *WACV*.

Nayak, G. K., Mopuri, K. R., Shaj, V., Babu, R. V. & Chakraborty, A. (2019). Zero-shot knowledge distillation in deep networks. In *ICML*.

Ng, R. W., Liu, X. & Swietojanski, P. (2018). Teacher-student training for text-independent speaker recognition. In *SLTW*.

Nie, X., Li, Y., Luo, L., Zhang, N. & Feng, J. (2019). Dynamic kernel distillation for efficient pose estimation in videos. In *ICCV*.

Noroozi, M., Vinjimoor, A., Favaro, P. & Pirsiavash, H. (2018). Boosting self-supervised learning via knowledge transfer. In *CVPR*.

Nowak, T. S. & Corso, J. J. (2018). Deep net triage: Analyzing the importance of network layers via structural compression. arXiv preprint arXiv:1801.04651.

Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., et al. (2018). Parallel wavenet: Fast high-fidelity speech synthesis. In *ICML*.

Pan, B., Cai, H., Huang, D. A., Lee, K. H., Gaidon, A., Adeli, E., & Niebles, J. C. (2020). Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*.

Pan, Y., He, F., & Yu, H. (2019). A novel enhanced collaborative autoencoder with knowledge distillation for top-n recommender systems. *Neurocomputing, 332,* 137–148.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I. & Talwar, K. (2017). Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*.

Papernot, N., McDaniel, P., Wu, X., Jha, S. & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE SP*.

Park, S. & Kwak, N. (2020). Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks. In *ECAI*.

Park, W., Kim, D., Lu, Y. & Cho, M. (2019). Relational knowledge distillation. In *CVPR*.

Passban, P., Wu, Y., Rezagholizadeh, M., & Liu, Q. (2021). ALP-KD: Attention-based layer projection for knowledge distillation. In *AAAI*.

Passalis, N. & Tefas, A. (2018). Learning deep representations with probabilistic knowledge transfer. In *ECCV*.

Passalis, N., Tzelepi, M., & Tefas, A. (2020a). Probabilistic knowledge transfer for lightweight deep representation learning. *TNNLS*. https://doi.org/10.1109/TNNLS.2020.2995884.

Passalis, N., Tzelepi, M., & Tefas, A. (2020b). Heterogeneous knowledge distillation using information flow modeling. In *CVPR*.

Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., et al. (2019a). Correlation congruence for knowledge distillation. In *ICCV*.

Peng, H., Du, H., Yu, H., Li, Q., Liao, J., & Fu, J. (2020). Cream of the crop: Distilling prioritized paths for one-shot neural architecture search. In *NeurIPS*.

Peng, Z., Li, Z., Zhang, J., Li, Y., Qi, G. J. & Tang, J. (2019b). Few-shot image recognition with knowledge transfer. In *ICCV*.

Perez, A., Sanguineti, V., Morerio, P. & Murino, V. (2020). Audio-visual model distillation using acoustic images. In *WACV*.

Phuong, M., & Lampert, C. H. (2019a). Towards understanding knowledge distillation. In *ICML*.

Phuong, M., & Lampert, C. H. (2019b). Distillation-based training for multi-exit architectures. In *ICCV*.

Pilzer, A., Lathuiliere, S., Sebe, N. & Ricci, E. (2019). Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *CVPR*.

Polino, A., Pascanu, R. & Alistarh, D. (2018). Model compression via distillation and quantization. In *ICLR*.

Price, R., Iso, K., & Shinoda, K. (2016). Wise teachers train better DNN acoustic models. *EURASIP Journal on Audio, Speech, and Music Processing, 2016*(1), 10.

Radosavovic, I., Dollar, P., Girshick, R., Gkioxari, G., & He, K. (2018). Data distillation: Towards omni-supervised learning. In *CVPR*.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollar P. (2020). Designing network design spaces. In *CVPR*.

Roheda, S., Riggan, B. S., Krim, H. & Dai, L. (2018). Cross-modality distillation: A case for conditional generative adversarial networks. In *ICASSP*.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). Fitnets: Hints for thin deep nets. In *ICLR*.

Ross, A. S. & Doshi-Velez, F. (2018). Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AAAI*.

Ruder, S., Ghaffari, P. & Breslin, J. G. (2017). Knowledge adaptation: Teaching to adapt. arXiv preprint arXiv:1702.02052.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*.

Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Saputra, M. R. U., de Gusmao, P. P., Almalioglu, Y., Markham, A. & Trigoni, N. (2019). Distilling knowledge from a deep pose regressor network. In *ICCV*.

Sau, B. B. & Balasubramanian, V. N. (2016). Deep model compression: Distilling knowledge from noisy teachers. arXiv preprint arXiv:1610.09650.

Seo, H., Park, J., Oh, S., Bennis, M., & Kim, S. L. (2020). Federated Knowledge Distillation. arXiv preprint arXiv:2011.02367.

Shakeri, S., Sethy, A. & Cheng, C. (2019). Knowledge distillation in document retrieval. arXiv preprint arXiv:1911.11065.

Shen, C., Wang, X., Song, J., Sun, L., & Song, M. (2019a). Amalgamating knowledge towards comprehensive classification. In *AAAI*.

Shen, C., Wang, X., Yin, Y., Song, J., Luo, S., & Song, M. (2021). Progressive network grafting for few-shot knowledge distillation. In *AAAI*.

Shen, C., Xue, M., Wang, X., Song, J., Sun, L., & Song, M. (2019b). Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation. In *ICCV*.

Shen, J., Vesdapunt, N., Boddeti, V. N. & Kitani, K. M. (2016). In teacher we trust: Learning compressed models for pedestrian detection. arXiv preprint arXiv:1612.00478.

Shen, P., Lu, X., Li, S. & Kawai, H. (2018). Feature representation of short utterances based on knowledge distillation for spoken language identification. In *Interspeech*.

Shen, P., Lu, X., Li, S., & Kawai, H. (2020). Knowledge distillation-based representation learning for short-utterance spoken language identification. *IEEE/ACM Transactions on Audio Speech and Language, 28,* 2674–2683.

Shen, P., Lu, X., Li, S. & Kawai, H. (2019c). Interactive learning of teacher-student model for short utterance spoken language identification. In *ICASSP*.

Shen, Z., He, Z. & Xue, X. (2019d). Meal: Multi-model ensemble via adversarial learning. In *AAAI*.

Shi, B., Sun, M., Kao, C. C., Rozgic, V., Matsoukas, S. & Wang, C. (2019a). Compression of acoustic event detection models with quantized distillation. In *Interspeech*.

Shi, B., Sun, M., Kao, CC., Rozgic, V., Matsoukas, S. & Wang, C. (2019b). Semi-supervised acoustic event detection based on tri-training. In *ICASSP*.

Shi, Y., Hwang, M. Y., Lei, X., & Sheng, H. (2019c). Knowledge distillation for recurrent neural network language modeling with trust regularization. In *ICASSP*.

Shin, S., Boo, Y. & Sung, W. (2019). Empirical analysis of knowledge distillation technique for optimization of quantized deep neural networks. arXiv preprint arXiv:1909.01688.

Shmelkov, K., Schmid, C., & Alahari, K. (2017). Incremental learning of object detectors without catastrophic forgetting. In *ICCV*.

Shu, C., Li, P., Xie, Y., Qu, Y., Dai, L., & Ma, L.(2019). Knowledge squeezed adversarial network compression. arXiv preprint arXiv:1904.05100.

Siam, M., Jiang, C., Lu, S., Petrich, L., Gamal, M., Elhoseiny, M., et al. (2019). Video object segmentation using teacher-student adaptation in a human robot interaction (HRI) setting. In *ICRA*.

Sindhwani, V., Sainath, T. & Kumar, S. (2015). Structured transforms for small-footprint deep learning. In *NeurIPS*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489.

Song, X., Feng, F., Han, X., Yang, X., Liu, W. & Nie, L. (2018). Neural compatibility modeling with attentive knowledge distillation. In *SIGIR*.

Srinivas, S. & Fleuret, F. (2018). Knowledge transfer with jacobian matching. In *ICML*.

Su, J. C. & Maji, S. (2017). Adapting models to signal degradation using distillation. In *BMVC*.

Sun, L., Gou, J., Yu, B., Du, L., & Tao, D. (2021) Collaborative teacher–student learning via multiple knowledge transfer. arXiv preprint arXiv:2101.08471.

Sun, S., Cheng, Y., Gan, Z. & Liu, J. (2019). Patient knowledge distillation for bert model compression. In *NEMNLP-IJCNLP*.

Sun, P., Feng, W., Han, R., Yan, S., & Wen, Y. (2019). Optimizing network performance for distributed dnn training on gpu clusters: Imagenet/alexnet training in 1.5 minutes. arXiv preprint arXiv:1902.06855.

Takashima, R., Li, S. & Kawai, H. (2018). An investigation of a knowledge distillation method for CTC acoustic models. In *ICASSP*.

Tan, H., Liu, X., Liu, M., Yin, B., & Li, X. (2021). KT-GAN: Knowledge-transfer generative adversarial network for text-to-image synthesis. *IEEE TIP, 30,* 1275–1290.

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., & Le, Q. V. (2019). Mnasnet: Platform-aware neural architecture search for mobile. In *CVPR*.

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*.

Tan, X., Ren, Y., He, D., Qin, T., Zhao, Z. & Liu, T. Y. (2019). Multi-lingual neural machine translation with knowledge distillation. In *ICLR*.

Tang, J., Shivanna, R., Zhao, Z., Lin, D., Singh, A., Chi, E. H., & Jain, S. (2020). Understanding and improving knowledge distillation. arXiv preprint arXiv:2002.03532.

Tang, J., & Wang, K. (2018). Ranking distillation: Learning compact ranking models with high performance for recommender system. In *SIGKDD*.

Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O. & Lin, J. (2019). Distilling task-specific knowledge from bert into simple neural networks. arXiv preprint arXiv:1903.12136.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*.

Thoker, F. M. & Gall, J. (2019). Cross-modal knowledge distillation for action recognition. In *ICIP*.

Tian, Y., Krishnan, D. & Isola, P. (2020). Contrastive representation distillation. In *ICLR*.

Tu, Z., He, F., & Tao, D. (2020). Understanding generalization in recurrent neural networks. In *International conference on learning representations*. *ICLR*.

Tung, F., & Mori, G. (2019). Similarity-preserving knowledge distillation. In *ICCV*.

Turc, I., Chang, M. W., Lee, K. & Toutanova, K. (2019). Well-read students learn better: The impact of student initialization on knowledge distillation. arXiv preprint arXiv:1908.08962.

Urban, G., Geras, K. J., Kahou, S. E., Aslan, O., Wang, S., Caruana, R., (2017). Do deep convolutional nets really need to be deep and convolutional? In *ICLR*.

Vapnik, V., & Izmailov, R. (2015). Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research, 16*(1), 2023–2049.

Vongkulbhisal, J., Vinayavekhin, P. & Visentini-Scarzanella, M. (2019). Unifying heterogeneous classifiers with distillation. In *CVPR*.

Walawalkar, D., Shen, Z., & Savvides, M. (2020). Online ensemble model compression using knowledge distillation. In *ECCV*.

Wang, C., Lan, X. & Zhang, Y. (2017). Model distillation with knowledge transfer from face classification to alignment and verification. arXiv preprint arXiv:1709.02929.

Wang, L., & Yoon, K. J. (2020). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. arXiv preprint arXiv:2004.05937.

Wang, H., Zhao, H., Li, X. & Tan, X. (2018a). Progressive blockwise knowledge distillation for neural network acceleration. In *IJCAI*.

Wang, J., Bao, W., Sun, L., Zhu, X., Cao, B., & Philip, S. Y. (2019a). Private model compression via knowledge distillation. In *AAAI*.

Wang, J., Gou, L., Zhang, W., Yang, H., & Shen, H. W. (2019b). Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *TVCG, 25*(6), 2168–2180.

Wang, M., Liu, R., Abe, N., Uchida, H., Matsunami, T., & Yamada, S. (2018b). Discover the effective strategy for face recognition model compression by improved knowledge distillation. In *ICIP*.

Wang, M., Liu, R., Hajime, N., Narishige, A., Uchida, H. & Matsunami, T.(2019c). Improved knowledge distillation for training fast low resolution face recognition model. In *ICCVW*.

Wang, T., Yuan, L., Zhang, X. & Feng, J. (2019d). Distilling object detectors with fine-grained feature imitation. In *CVPR*.

Wang, T., Zhu, J. Y., Torralba, A., & Efros, A. A. (2018c). Dataset distillation. arXiv preprint arXiv:1811.10959.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020a). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.

Wang, W., Zhang, J., Zhang, H., Hwang, M. Y., Zong, C. & Li, Z. (2018d). A teacher-student framework for maintainable dialog manager. In *EMNLP*.

Wang, X., Fu, T., Liao, S., Wang, S., Lei, Z., & Mei, T. (2020b). Exclusivity-consistency regularized knowledge distillation for face recognition. In *ECCV*.

Wang, X., Hu, J. F., Lai, J. H., Zhang, J. & Zheng, W. S. (2019e). Progressive teacher-student learning for early action prediction. In *CVPR*.

Wang, X., Zhang, R., Sun, Y. & Qi, J. (2018e) Kdgan: Knowledge distillation with generative adversarial networks. In *NeurIPS*.

Wang, Y., Xu, C., Xu, C., & Tao, D. (2019f). Packing convolutional neural networks in the frequency domain. *IEEE TPAMI, 41*(10), 2495–2510.

Wang, Y., Xu, C., Xu, C. & Tao, D. (2018f). Adversarial learning of portable student networks. In *AAAI*.

Wang, Z. R., & Du, J. (2021). Joint architecture and knowledge distillation in CNN for Chinese text recognition. *Pattern Recognition, 111,* 107722.

Watanabe, S., Hori, T., Le Roux, J. & Hershey, J. R. (2017). Student-teacher network learning with enhanced features. In *ICASSP*.

Wei, H. R., Huang, S., Wang, R., Dai, X. & Chen, J. (2019). Online distilling from checkpoints for neural machine translation. In *NAACL-HLT*.

Wei, Y., Pan, X., Qin, H., Ouyang, W. & Yan, J. (2018). Quantization mimic: Towards very tiny CNN for object detection. In *ECCV*.

Wong, J. H. & Gales, M. (2016). Sequence student-teacher training of deep neural networks. In *Interspeech*.

Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., et al. (2019). Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *CVPR*.

Wu, A., Zheng, W. S., Guo, X. & Lai, J. H. (2019a). Distilled person re-identification: Towards a more scalable system. In *CVPR*.

Wu, G., & Gong, S. (2021). Peer collaborative learning for online knowledge distillation. In *AAAI*.

Wu, J., Leng, C., Wang, Y., Hu, Q. & Cheng, J. (2016). Quantized convolutional neural networks for mobile devices. In *CVPR*.

Wu, M. C., Chiu, C. T. & Wu, K. H. (2019b). Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks. In *ICASSP*.

Wu, X., He, R., Hu, Y., & Sun, Z. (2020). Learning an evolutionary embedding via massive knowledge distillation. *International Journal of Computer Vision,* 1–18.

Xia, S., Wang, G., Chen, Z., & Duan, Y. (2018). Complete random forest based class noise filtering learning for improving the generalizability of classifiers. *IEEE TKDE, 31*(11), 2063–2078.

Xie, J., Lin, S., Zhang, Y. & Luo, L. (2019). Training convolutional neural networks with cheap convolutions and online distillation. arXiv preprint arXiv:1909.13063.

Xie, Q., Hovy, E., Luong, M. T., & Le, Q. V. (2020). Self-training with Noisy Student improves ImageNet classification. In *CVPR*.

Xu, G., Liu, Z., Li, X., & Loy, C. C. (2020a). Knowledge distillation meets self-supervision. In *ECCV*.

Xu, K., Rui, L., Li, Y., & Gu, L. (2020b). Feature normalized knowledge distillation for image classification. In *ECCV*.

Xu, Z., Wu, K., Che, Z., Tang, J., & Ye, J. (2020c). Knowledge transfer in multi-task deep reinforcement learning for continuous control. In *NeurIPS*.

Xu, Z., Hsu, Y. C. & Huang, J. (2018a). Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. In *ICLR workshop*.

Xu, Z., Hsu, Y. C. & Huang, J. (2018b). Training student networks for acceleration with conditional adversarial networks. In *BMVC*.

Xu, T. B., & Liu, C. L. (2019). Data-distortion guided self-distillation for deep neural networks. In *AAAI*.

Yan, M., Zhao, M., Xu, Z., Zhang, Q., Wang, G. & Su, Z. (2019). Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In *ICCVW*.

Yang, C., Xie, L., Qiao, S. & Yuille, A. (2019a). Knowledge distillation in generations: More tolerant teachers educate better students. In *AAAI*.

Yang, C., Xie, L., Su, C. & Yuille, A. L. (2019b). Snapshot distillation: Teacher-student optimization in one generation. In *CVPR*.

Yang, J., Martinez, B., Bulat, A., & Tzimiropoulos, G. (2020a). Knowledge distillation via adaptive instance normalization. In *ECCV*.

Yang, Y., Qiu, J., Song, M., Tao, D. & Wang, X. (2020b). Distilling knowledge from graph convolutional networks. In *CVPR*.

Yang, Z., Shou, L., Gong, M., Lin, W. & Jiang, D. (2020c). Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *WSDM*.

Yao, A., & Sun, D. (2020). Knowledge transfer via dense cross-layer mutual-distillation. In *ECCV*.

Yao, H., Zhang, C., Wei, Y., Jiang, M., Wang, S., Huang, J., Chawla, N. V., & Li, Z. (2020). Graph few-shot learning via knowledge transfer. In *AAAI*.

Ye, J., Ji, Y., Wang, X., Gao, X., & Song, M. (2020). Data-free knowledge amalgamation via group-stack dual-GAN. In *CVPR*.

Ye, J., Ji, Y., Wang, X., Ou, K., Tao, D. & Song, M. (2019). Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In *CVPR*.

Yim, J., Joo, D., Bae, J. & Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*.

Yin, H., Molchanov, P., Alvarez, J. M., Li, Z., Mallya, A., Hoiem, D., Jha, Niraj K., & Kautz, J. (2020). Dreaming to distill: Data-free knowledge transfer via DeepInversion. In *CVPR*.

Yoo, J., Cho, M., Kim, T., & Kang, U. (2019). Knowledge extraction with no observable data. In *NeurIPS*.

You, S., Xu, C., Xu, C., & Tao, D. (2017). Learning from multiple teacher networks. In *SIGKDD*.

You, S., Xu, C., Xu, C. & Tao, D. (2018). Learning with single-teacher multi-student. In *AAAI*.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., et al. (2019). Large batch optimization for deep learning: Training bert in 76 minutes. In *ICLR*.

Yu, L., Yazici, V. O., Liu, X., Weijer, J., Cheng, Y. & Ramisa, A. (2019). Learning metrics from teachers: Compact networks for image embedding. In *CVPR*.

Yu, X., Liu, T., Wang, X., & Tao, D. (2017). On compressing deep models by low rank and sparse decomposition. In *CVPR*.

Yuan, F., Shou, L., Pei, J., Lin, W., Gong, M., Fu, Y., & Jiang, D. (2021). Reinforced multi-teacher selection for knowledge distillation. In *AAAI*.

Yuan, L., Tay, F. E., Li, G., Wang, T. & Feng, J. (2020). Revisit knowledge distillation: a teacher-free framework. In *CVPR*.

Yuan, M., & Peng, Y. (2020). CKD: Cross-task knowledge distillation for text-to-image synthesis. *IEEE TMM, 22*(8), 1955–1968.

Yue, K., Deng, J., & Zhou, F. (2020). Matching guided distillation. In *ECCV*.

Yun, S., Park, J., Lee, K. & Shin, J. (2020). Regularizing class-wise predictions via self-knowledge distillation. In *CVPR*.

Zagoruyko, S. & Komodakis, N. (2017). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*.

Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M. & Mori, G. (2019). Lifelong gan: Continual learning for conditional image generation. In *ICCV*.

Zhai, S., Cheng, Y., Zhang, Z. M. & Lu, W. (2016). Doubly convolutional neural networks. In *NeurIPS*.

Zhao, C., & Hospedales, T. (2020). Robust domain randomised reinforcement learning through peer-to-peer distillation. In *NeurIPS*.

Zhao, H., Sun, X., Dong, J., Chen, C., & Dong, Z. (2020a). Highlight every step: Knowledge distillation via collaborative teaching. *IEEE TCYB*. https://doi.org/10.1109/TCYB.2020.3007506.

Zhao, L., Peng, X., Chen, Y., Kapadia, M., & Metaxas, D. N. (2020b). Knowledge as Priors: Cross-Modal Knowledge Generalization for Datasets without Superior Knowledge. In *CVPR*.

Zhao, M., Li, T., Abu Alsheikh, M., Tian, Y., Zhao, H., Torralba, A. & Katabi, D. (2018). Through-wall human pose estimation using radio signals. In *CVPR*.

Zhang, C. & Peng, Y. (2018). Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. In *IJCAI*.

Zhang, F., Zhu, X. & Ye, M. (2019a). Fast human pose estimation. In *CVPR*.

Zhang, J., Liu, T., & Tao, D. (2018). An information-theoretic view for deep learning. arXiv preprint arXiv:1804.09060.

Zhang, H., Hu, Z., Qin, W., Xu, M., & Wang, M. (2021a). Adversarial co-distillation learning for image recognition. *Pattern Recognition, 111*, 107659.

Zhang, L., Shi, Y., Shi, Z., Ma, K., & Bao, C. (2020a). Task-oriented feature distillation. In *NeurIPS*.

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C. & Ma, K. (2019b). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*.

Zhang, M., Song, G., Zhou, H., & Liu, Y. (2020b). Discriminability distillation in group representation learning. In *ECCV*.

Zhang, S., Feng, Y., & Li, L. (2021b). Future-guided incremental transformer for simultaneous translation. In *AAAI*.

Zhang, S., Guo, S., Wang, L., Huang, W., & Scott, M. R. (2020c). Knowledge integration networks for action recognition. In *AAAI*.

Zhang, W., Miao, X., Shao, Y., Jiang, J., Chen, L., Ruas, O., & Cui, B. (2020d). Reliable data distillation on graph convolutional network. In *ACM SIGMOD*.

Zhang, X., Wang, X., Bian, J. W., Shen, C., & You, M. (2021c). Diverse knowledge distillation for end-to-end person search. In *AAAI*.

Zhang, X., Zhou, X., Lin, M. & Sun, J. (2018a). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*.

Zhang, Y., Lan, Z., Dai, Y., Zeng, F., Bai, Y., Chang, J., & Wei, Y. (2020e). Prime-aware adaptive distillation. In *ECCV*.

Zhang, Y., Xiang, T., Hospedales, T. M. & Lu, H. (2018b). Deep mutual learning. In *CVPR*.

Zhang, Z., & Sabuncu, M. R. (2020). Self-distillation as instance-specific label smoothing. In *NeurIPS*.

Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., & Zha, Z. J. (2020f). Object relational graph with teacher-recommended learning for video captioning. In *CVPR*.

Zhou C, Neubig G, Gu J (2019a) Understanding knowledge distillation in non-autoregressive machine translation. In *ICLR*.

Zhou, G., Fan, Y., Cui, R., Bian, W., Zhu, X. & Gai, K. (2018). Rocket launching: A universal and efficient framework for training well-performing light net. In *AAAI*.

Zhou, J., Zeng, S. & Zhang, B. (2019b) Two-stage image classification supervised by a single teacher single student model. In *BMVC*.

Zhou, P., Mai, L., Zhang, J., Xu, N., Wu, Z. & Davis, L. S. (2020). M2KD: Multi-model and multi-level knowledge distillation for incremental learning. In *BMVC*.

Zhu, M., Han, K., Zhang, C., Lin, J., & Wang, Y. (2019). Low-resolution visual recognition via deep feature distillation. In *ICASSP*.

Zhu, X., & Gong, S. (2018). Knowledge distillation by on-the-fly native ensemble. In *NeurIPS*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.