MAKD:MULTIPLE AUXILIARY KNOWLEDGE DISTILLATION

Zehan Chen, Xuan Jin, Yuan He, Hui Xue

Alibaba Group

ABSTRACT

Knowledge distillation aims to learn a small student model by leveraging knowledge from a larger teacher model. The gap between these heterogeneous models hinder their knowledge transfer and it would be more challenging when the teacher model is from another task. Previous methods view the teacher model as a perfect feature extractor and train the student model to mimic it. While we notice that the teacher model has defect in extracting features of another task samples. To improve knowledge distillation under such situation, we propose Multiple Auxiliary Subspaces (MAS). Most previous methods improve distillation performance by representation alignment, while we resort to the promotion of the teacher model which is more suitable for cross-task distillation. The MAS distills the knowledge in a mutual learning way based on an auxiliary network. Along with the training procedure, the teacher model is improved by the auxiliary network which works as a trainable part of the teacher model and learn the features distribution of target samples from the student model. And this promotion of the teacher model will benefit the student model via the following distillation procedure. We adopt the representation alignment technique, multiple auxiliary networks to further enhance the proposed method. The MAS works well with limited or sufficient labeled target data. If the source data is available, based on it the MAS can construct another auxiliary network to further improve the distillation. Experiments are conducted to validate that the MAS outperforms baseline methods and achieves state-of-the-art results on several standard benchmarks.

***Index Terms*—** Knowledge distillation, cross-task, auxiliary network

1. INTRODUCTION

Deep Neural Networks (DNN) have achieved great success in various tasks, driven by large-scale labeled data and complex network structure. Due to the time-consuming labeling procedure, deployment and computational consumption issues, sometimes we are required to train a small DNN with limited labeled data. Knowledge distillation method [1] allow us to learn a small student model by leveraging knowledge from a larger teacher model. Getting a teacher model by training a large model or construct an ensemble model with ensemble learning method [2] is time consuming. As there are plenty of available pre-trained models in the Internet which is trained with large-scale labeled data such as ImageNet [3]. Those models can be used as teacher model if we conduct the knowledge distillation in a suitable way.

In this paper, we propose a method to distill the knowledge from a large teacher model to a smaller student model, while the teacher model and student model are not in the same task. We note the teacher model and its training data as source model and source data respectively. And the student model and its training data is noted as target model and target data respectively. By leveraging the knowledge of a heterogeneous source model, we aim to promote the performance of a small target model that trained with limited labeled target data, with or without the help of source data. In some situation, we may tend to obtain sufficient annotation target data when we make trade off between annotation cost and target model performance. So we consider this in our proposed method to distill knowledge with sufficient target data.

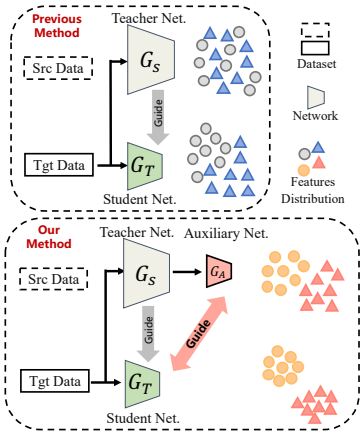

Fig. 1. The proposed auxiliary network reduces the output feature's distribution gap between teacher network and student network to improve the knowledge distillation.

In knowledge distillation methods [1, 4, 5] the student model is trained to mimic the outputs of samples represented by the teacher model. While the gap between teacher (source) model and student (target) model would hinder their knowledge transfer, especially when the two models perform on different tasks. To reducing the gap, representation alignment is adopted in previous method like [5, 6, 7], in which the parameter of source model is frozen and the feature selection and feature alignment manner are consider to improve the knowledge distillation. But the source model is under-fitting in target data, which may lead to sub-optimal guidance.

Motivated by this, as shown in Fig.1, the proposed method construct a light auxiliary network to further reduce the gap between source model and target model. To alleviate the alignment impact brought by distribution deviation, existing methods leverage the subspace and manifold learning [8, 9], to progressively transfer knowledge to the target model [10, 11]. Similarly we construct auxiliary network to adapting the source model's outputs to target data distribution and provide supplementary knowledge to source model. Specifically, the auxiliary network maps features between source model and target model and learns the mapping function. The distribution of target model would change while training, therefore the

proposed method utilizes a training scheme to empower the auxiliary network to keep updating its mapping capability.

To further reduce the gap by representation alignment [5, 6, 7], the proposed method is enhanced to produce multiple auxiliary subspaces. If the source data is available, the proposed method can construct another auxiliary network to further improve the distillation. Experiments are conducted to validate that the proposed method outperforms competitors by a large margin and achieves state-of-the-art results on several standard benchmarks. Our contributions are:

- A cross-task knowledge distillation method that enables mutual learning based on an auxiliary network.
- Multiple subspaces, multiple auxiliary networks are applied to formulate the framework of the proposed method.
- Applications to distillation with limited or sufficient target data. The proposed method achieves state-of-the-art results on several standard benchmarks.

2. METHOD

In this section, we first introduce a basic method of auxiliary subspace to help the representation alignment. Then, we enhance the method by multiple auxiliary subspaces. After that, to further utilize available source data, we proposed multiple auxiliary networks.

2.1. Auxiliary Subspace

In this paper, we investigate knowledge distillation, where the teacher (source) model is trained with data that different from the student (target) model's. We are given a pre-trained source model G_S, along with a target model G_T and a target classifier F_T to be optimized. There are N labeled target samples $\{x_i, y_i\}_{i=1}^N$ where $y_i \in \{1, ..., C_T\}$. The C_T categories of target task are different with the categories of source task. In practice, we may face different situations in knowledge distillation, where the labeled target data is limited or sufficient. Both of them are considered in this paper.

The goal of the proposed method is to obtain well-performed results on G_T with assistance of the source model G_S. The $G_S(x_i)$ could be regarded as source features. $G_S(x_i)$ and $G_T(x_i)$ could be regarded as paired features for further alignment. Under different network structures and distribution deviation, it is difficult to achieve high performance of G_T by direct alignment between outputs of G_S and G_T.

Following the subspace and manifold learning [8, 9], we construct a new auxiliary subspace, to improve the source features for more effective distribution alignment. As the parameters of the source model are frozen, we use an auxiliary network G_A to map the source features to the auxiliary subspace. The inputs of G_A are $G_S(x_i)$, while the outputs of G_A are in the same dimension with the outputs of $G_T(x_i)$. Then we align $G_A(G_S(x_i))$ to $G_T(x_i)$, in order to reduce the distribution gap between auxiliary subspace and target features. We construct auxiliary network by one single Fully-Connected layer in this paper. Meanwhile, one single Sequeze-and-Excitation block [12] can get excellent performance as well.

As the distribution of output features from $G_T(x_i)$ is changing during training procedure, we design a training scheme to empower the auxiliary network to synchronize with the target model to improve its capabilities. In the training scheme, there are three steps to iteratively update G_A and G_T, as shown in Fig 2 (c). In step 1, we simply train the target model G_T and target classifier F_T by classification loss. This step sets up the representation of target data for auxiliary network to leverage from. The classification loss is calculated by the cross-entropy loss on labeled target data, as follows:

$$L_{CLS} = \frac{1}{N}\sum^{N} L_{CE}(F_T(G_T(x_i)), y_i), \tag{1}$$

where L_{CE} denotes the cross-entropy loss. In the step 2, we train G_A by aligning the features of auxiliary network and target network with the help of alignment loss. In this step the G_A obtains knowledge from the target data, as shown in Fig 2 (a). Following previous method[13], we align the features by calculating mean square error (MSE) between $G_T(x_i)$ and $G_A(G_S(x_i))$, which can be summarized as follows:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(G_T(x_i) - G_A(G_S(x_i)))^2. \tag{2}$$

In the step 3, the parameters of G_A and G_S are frozen, and the parameters of G_T and F_T are updated with classification loss and matching loss. In this step the knowledge is transferred from G_A and G_S to G_T. The classification loss is the same with Eqn. 1 and the matching loss is the same as Eqn. 2. The step 2 and step 3 will be repeated untill the end of the training procedure.

These three steps formulate the basic training procedure of MAS, and gradually distill the knowledge to the target model. The step 1 could obtain basic parameters of G_T and F_T. the step 2 could obtain well-performed G_A by the alignment procedure. Actually, the step 2 is repeated several times to obtain well-performed G_A. Finally the Step 3 modifies the outputs of G_T with assistance of the outputs of G_S and G_A. The transferable knowledge could be dug out from G_S, to the subspace generated by G_A. During the training process, G_A could transfer the outputs of source model to the auxiliary subspace, with less discrepancy compared with the target features.

2.2. Multiple Auxiliary Subspaces

It has been concluded that the connection of multiple layer outputs between source and target model should be differently weighted for better knowledge transfer[7]. The weights could be learned in the automatic training procedure of the network. The auxiliary subspace could be further enhanced to Multiple Auxiliary Subspaces by multiple connections. As shown in Fig 3, the alignment could be achieved by one-to-one, one-to-multi and multi-to-multi matching.

We achieve multiple auxiliary subspaces by one-to-multi and multi-to-multi matching. To analyze the multiple auxiliary subspaces, we denote p and q as the layer number, with totally P and Q selected layers. In this section, we denote G_S^p and G_T^q as the source or target model from the beginning to the pth or qth layer. And we denote $F_{G_S} = \sum_{p=1}^{P} G_S^p(x_i)$ and $F_{G_T} = \sum_{q=1}^{Q} G_T^q(x_i)$ as the set of output features of each layer from the source or target model. The alignment between these features can be found in Fig 3. To guide the full target model, we leverage the one-to-multi matching by the outputs of G_S^P and each layer outputs of G_T^q. Then Eqn. 2 could be replaced as follows:

$$L_{MAS} = \frac{1}{NQ}\sum_{i=1}^{N}\sum_{q=1}^{Q}(G_N^{P,q}(G_A(G_S^P(x_i))) - G_T^q(x_i))^2, \tag{3}$$

where $G_N^{p,q}$ denotes the network to achieve the same dimension between source and target features of layer p and q. Similar to [7], $G_N^{p,q}$ consists of a bilinear interpolation layer and a convolutional

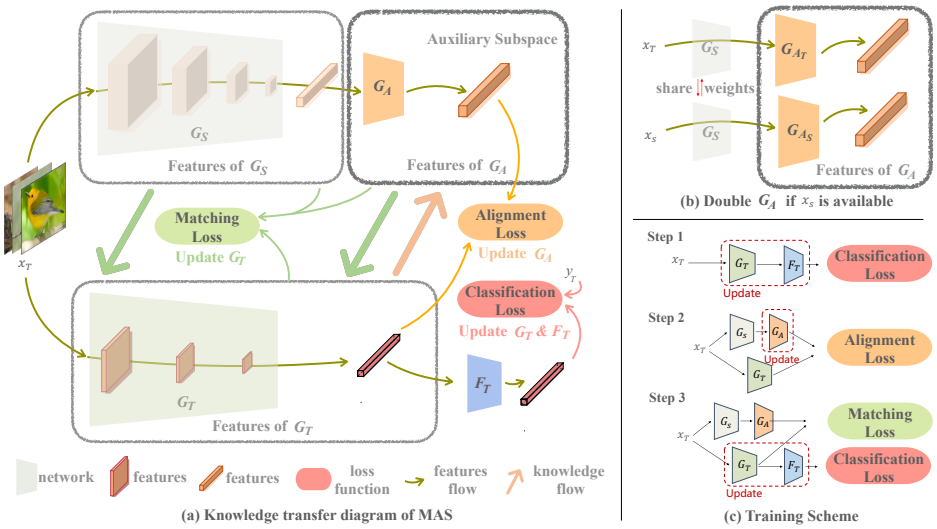

Fig. 2. Knowledge distillation diagram of the proposed method. We propose the auxiliary network(G_A) to estimate the data distribution of the target data to help the representation alignment. If the unlabeled source data is available, the proposed method generates another auxiliary network to trained on it and provides more knowledge to help representation alignment. With our training scheme, the auxiliary network synchronizes with the target model to improve its capabilities.

layer with the kernel size of 1×1. The outputs of $G_N^{P,q}$ could represent multiple auxiliary subspaces with different values of q, thus the method is denoted as MAS. In one-to-multi matching manner, P is a constant number, and the G_A only takes the last layer outputs of G_S.

Rich semantic features from different layers of the source model can be used as knowledge. Accordingly, we develop multi-to-multi matching, by alignment from all selected outputs in G_S^p to G_T^q. Then the alignment process could be calculated as follows:

$$L_{P2Q} = \frac{1}{NPQ} \sum_{i=1}^{N} \sum_{p=1}^{P} \sum_{q=1}^{Q} (G_N^{p,q}(G_S^p(x_i)) - G_T^q(x_i))^2, \quad (4)$$

$$L_{M^2AS} = L_{P2Q} + L_{MAS}, \quad (5)$$

The outputs of $G_N^{p,q}$ from 5 could represent multiple to multiple auxiliary subspaces (M^2AS). The only difference between AS, MAS, and M^2AS lies in the different alignment manner as shown in Fig 3. Similar to [7], the connection weights between source and target model are learned in the automatic training procedure of a meta network. The matching loss function of AS, as shown in Eqn. 2, is replaced by Eqn. 3 and Eqn. 5 to achieve MAS and M^2AS. When the latent features of source model is not available, we distill knowledge by AS or MAS, otherwise we choose M^2AS for better results.

2.3. Multiple Auxiliary Networks

In some situations the source data is available. Our method can generate another auxiliary network to make use of the source data and provide more transferable knowledge to further improve the distillation performance. And we denote this as M^2A^2S. In this case, G_{A_T} and G_{A_S} are two auxiliary networks and they are trained with different task data, as shown in Fig 2. Base on target data, the G_{A_T} maps the chaotic distribution of source features to orderly distribution. Base on source data, the G_{A_S} maps the orderly distribution of source features to chaotic distribution. These two opposite mapping functions, G_{A_T} and G_{A_S}, both provide information about the relationship between G_S and G_T. To achieve this, the step 2 of the training scheme trains two auxiliary networks independently while the step 1 and step 3 remain the same.

3. EXPERIMENTS

We evaluate the proposed method with limited or sufficient labeled target data in visual classification tasks.

3.1. Experimental Settings

Under limited data setting, CUB200 [14], Stanford40 [15] and MIT67 [16] are selected as target dataset which provide less than 6000 training samples. The ImageNet [3] is used as source dataset. Under the setting of distill with source data, we randomly select 1% training samples from ImageNet without annotation. While under sufficient data setting, three subsets of DomainNet [17], the clipart, the painting and the real are selected as source datasets. And they provide 33525, 50416 and 120906 training samples respectively. The painting subset is selected as target dataset. The Resnet-34 and the Resnet-18 [18] are used as source model and target model respectively. Several excellent methods, AT [6], FM [5], L2T [7] and CRD [19] are selected for comparison.

3.2. Distillation with Limited Target Data

We conduct the experiments with single auxiliary subspace in one-to-one matching setting, denoted as AS. We build up multiple auxiliary subspaces to distill knowledge in one-to-multi and multi-to-multi matching setting, as introduced in 2.2, which is denoted as MAS and M^2AS respectively. While the source data is available, we construct another auxiliary network to enrich the multiple auxiliary subspaces and denote it as M^2A^2S. The AT and FM connect

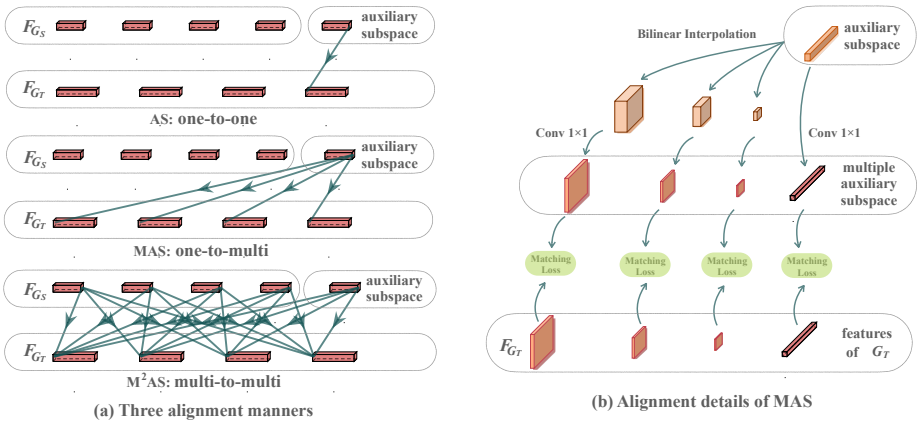

Fig. 3. The proposed MAS is flexible and has three knowledge distillation manners. In AS, it produces auxiliary subspace and distill directly. In MAS and M^{2}AS, it produces multiple auxiliary subspaces by one-to-multi and multi-to-multi matching.

Table 1. Accuracies (%) on CUB200, Stanford40 and MIT67. Source model is Resnet-34 and target model is Resnet-18.

Method	CUB200	Stanford40	MIT67
Scratch	42.15 ±0.45	36.93 ±0.56	48.91 ±1.89
Finetune	54.54 ±0.86	42.91 ±0.46	53.06 ±0.76
AT	57.74 ±1.17	59.29 ±0.91	59.18 ±1.57
FM	48.93 ±0.40	44.50 ±0.96	54.88 ±1.24
L2T	65.05 ±1.19	63.08 ±0.88	64.85 ±2.75
CRD	67.58 ±2.27	66.14 ±1.69	55.65 ±0.24
AS	69.53 ±0.47	66.71 ±0.31	66.61 ±0.48
MAS	70.28 ±0.91	68.85 ±0.17	68.66 ±1.34
M^{2}AS	**73.74 ±0.75**	**73.88 ±0.40**	**72.49 ±0.57**
M^{2}A^{2}S	**74.52 ±0.54**	**74.90 ±0.30**	**73.25 ±0.35**

each feature layer before pooling in source model to the counterpart in target model. In L2T, except the first convolutional layer, each feature layer before pooling in source model is connected to each feature layer before pooling in target model.

The proposed method AS outperforms all baseline methods in all three datasets as shown in Table 1. Comparing to the results of finetuning and training from scratch, our method can successfully transfer valuable knowledge from auxiliary subspace to the target model, and there are less negative transfer in our proposed method.The baseline methods transfer knowledge with effective feature matching technique while AS just simply transfer the knowledge from auxiliary subspace. But the results still show the superiority of AS. This indicates that it is valuable to adapt source features to target distribution before knowledge distillation.

As shown in Table 1, both MAS and M^{2}AS get better performance than AS, as well as all the baseline methods. Comparing baseline methods, M^{2}AS improves the accuracy by 6.16%, 7.74%, 7.64% on CUB200, Stanford40 and MIT67 datasets respectively. Comparing to the AS, the performance of the target models has been greatly improved in all the three datasets when we use multiple auxiliary subspaces. This validates the effectiveness of our method, and the knowledge of the auxiliary subspace and source model can complement each other to jointly improve the performance of the target model. With additional 1% of unlabeled source images, the M^{2}A^{2}S can further improves the classification accuracy.

Table 2. Accuracies (%) on DomainNet.

Method	Clipart	Real	Painting
Scratch	61.02 ±0.52	61.02 ±0.73	61.02 ±0.44
Finetune	60.47 ±0.68	60.47 ±0.51	60.47 ±0.33
L2T	57.21 ±1.21	58.79 ±0.99	57.02 ±0.35
CRD	61.42 ±0.84	61.78 ±1.23	61.59 ±0.84
M^{2}AS	**61.46 ±0.22**	**62.03 ±0.58**	**61.89 ±0.74**
M^{2}A^{2}S	**61.77 ±0.21**	**62.16 ±0.66**	**62.37 ±0.52**

3.3. Distillation with Sufficient Labeled Target Data

The performance of the finetune shown in Table 2 is lower than the scratch, which indicates that initializing the target model with pre-trained parameters from ImageNet [3] may cause minor negative effects. Because the target task is quite different from the task in the ImageNet, and the labeled data in target task is sufficient enough to train a high performance target model. The L2T in Table 2 gets lower performance than the scratch. We infer that distillation would be more challenge with sufficient data.

As shown in Table 2, experiments of M^{2}AS and M^{2}A^{2}S perform better than the comparing methods. The proposed method makes trade off between knowledge from target data and teacher model as previous methods. While our auxiliary network, that maps features between teacher model and student model, makes contribution by emphasizing the knowledge from target data.

4. CONCLUSION

In this paper, we consider the knowledge distillation where the teacher model and student model are performed on different tasks. Different from previous methods, we improve the knowledge distillation by the promotion of the teacher model, and distill the knowledge in a mutual learning way based on an auxiliary network. The proposed method achieves state-of-the-art results with limited or sufficient target labeled data.

5. REFERENCES

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[2] Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang, "Agree to disagree: Adaptive ensemble knowledge distillation in gradient space," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[3] Jia Deng, Alex Berg, Sanjeev Satheesh, H Su, Aditya Khosla, and L Fei-Fei, "Imagenet large scale visual recognition competition 2012 (ilsvrc2012)," *See net. org/challenges/LSVRC*, p. 41, 2012.

[4] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

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

[7] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin, "Learning what and where to transfer," *arXiv preprint arXiv:1905.05901*, 2019.

[8] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2066–2073.

[9] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2960–2967.

[10] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool, "Dlow: Domain flow for adaptation and generalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2477–2486.

[11] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian, "Gradually vanishing bridge for adversarial domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12455–12464.

[12] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[13] Zhizhong Li and Derek Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.

[15] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *2011 International conference on computer vision*. IEEE, 2011, pp. 1331–1338.

[16] Ariadna Quattoni and Antonio Torralba, "Recognizing indoor scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 413–420.

[17] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1406–1415.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] Yonglong Tian, Dilip Krishnan, and Phillip Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.