Data-Free Knowledge Distillation For Image Super-Resolution Based On Contrastive Learning

Pei Wang[1], **Shenghao Nie**[1], **Jianxiang Xie**[1], **Yifan Wang**[2], **Chaoxing Zhang**[1]

[1]Artificial Intelligence, [2]School of Informatic

{23020231154229, 23020231154217, 23020231154237, 31520231154321, 36120221150469}@stu.xmu.edu.cn

Abstract

Data-free Knowledge Distillation allows learning the student network from the teacher network without the need for training data, thereby making the model more lightweight while reducing the model computation and memory requirements, all while maintaining relatively high performance. In the context of image super-resolution, the data-free knowledge distillation method faces difficulties in ensuring the diversity of training data generated within the same training batch and across different batches. Thus, we modified the data generation approach of the DFSR generator, introducing Contrastive Memory Inversion (CMI) to enhance the differences between the samples generated in each instance and the historical samples in the memory bank, thereby increasing sample diversity. Experimental results demonstrate that our proposed network can achieve better results than DFSR for both quantitative and qualitative results.

Introduction

Deep learning has achieved significant success in numerous domains, including image recognition (Pak and Kim 2017), natural language processing (Chowdhary and Chowdhary 2020), and image super-resolution (Yang et al. 2019). Over the past few decades, images have become a crucial medium for transmitting information over the internet. While the performance of various hardware has improved significantly, many small devices are unable to accommodate high-performance hardware. Yet, these devices are expected to display clear images by obtaining high-resolution images from low-resolution cameras. Moreover, the bandwidth costs associated with image transmission may be substantial, which can be reduced effectively by image compression technology. Therefore, image super-resolution(ISR) has come into people's focus.

The underlying principle of ISR (Hunt 1995) based on deep learning involves training a neural network using a dataset comprising a large number of low-resolution and high-resolution images. This training allows the neural network to learn the mapping relationship from low to high resolution. When a pre-trained image super-resolution model is deployed on hardware-constrained devices, it allows for obtaining high-quality images at a relatively low cost.

Numerous successful works (Dai et al. 2019; Niu et al. 2020; Wang et al. 2022) have been developed based on the aforementioned principle. However, these models are becoming increasingly complex and demanding in terms of computability. Consequently, there is a growing demand for structurally simplified yet high-performing models. To achieve better models, knowledge distillation (Zhao et al. 2022) methods are being explored. The objective of knowledge distillation is to extract concise and high-performance student models from extensive pre-trained teacher models, which have high requirements for datasets and often perform poorly in the absence of a suitable dataset due to some privacy or transmission constraints. Data-free training based on generative models can address this issue. However, generative networks commonly suffer from mode collapse, where the generated instances turn out to be highly similar to each other.

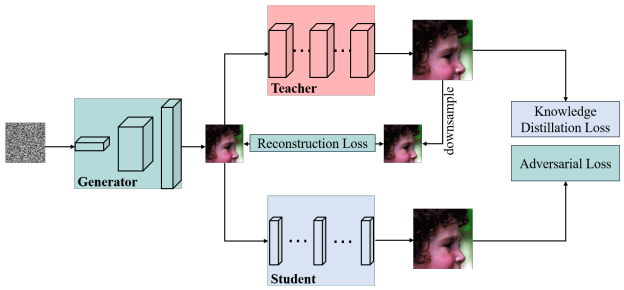

Figure 1: Framework of data-free knowledge distillation for image super-resolution.

To mitigate this mode collapse issue, we introduce the contrastive learning approach into data-free distillation methods for image super-resolution tasks to enhance the diversity of training samples. Our method consists of two components: image generation and knowledge distillation. Unlike traditional data-free knowledge distillation where the entire training samples for an epoch are synthesized at once, We adopt a contrastive model inversion (CMI) approach to progressively synthesize new samples that can be easily distinguished from historical samples in the memory bank. The generator network generates a low-resolution image based on noise, while the student network can learn knowledge from the teacher network through distillation loss with the guidance of the high-resolution images provided by the

teacher network. To further improve the diversity of the generated data, we propose to use adversarial loss to optimize the generator to maximize the difference between the student network and the teacher network. Besides, a reconstruction loss is introduced to optimize the generator to make the generated images closer to the original data distribution by constraining the downsampled output from the teacher network to be consistent with the low-resolution images generated by the generator. Experimental results demonstrate that our proposed network can achieve good results even without the original training data.

Related Work

Image Super-Resolution (SR) aims to generate a high-resolution image from the low-resolution version. While there are urgent demands for applying image superresolution networks to mobile devices such as cellphones and cameras, recently, efficient and lightweight SR networks have attracted increasing interest in the computer vision community. Knowledge distillation is a common method for obtaining lightweight models, and researchers have made many attempts in this regard.

The work (Gao et al. 2019) involves generating various statistical representations from feature maps as a means of extracting valuable information from teacher super-resolution networks. (He et al. 2020b) introduce a method called Feature Affinity-Based Knowledge Distillation (FAKD) for enhancing the distillation performance of super-resolution networks. This approach leverages feature map correlations to improve the knowledge transfer process. In addition, (Hui, Wang, and Gao 2018) and (Jiang et al. 2018) design new structures to perform distillation between different parts of the model and improve the performance of the lightweight super-resolution network.

The methods mentioned earlier produce efficient models that deliver strong performance while demanding minimal computational resources. Nevertheless, implementing these techniques typically necessitates access to the original training dataset. In real-world scenarios, this dataset is frequently unavailable due to privacy or data transfer considerations. Consequently, it becomes crucial to investigate data-free model compression approaches.

Data-Free Knowledge Distillation geared to learn a portable network without any training data. The primary task of these methods is to obtain training data. (Nayak et al. 2019) represent the teacher network's output as a Dirichlet distribution and repeatedly manipulate input noisy images to generate a set of training samples. In contrast to the iterative optimization of noise images, (Chen et al. 2019) utilize a Generative Adversarial Network (GAN) to create training samples. They achieve this by adjusting the generator network's parameters using a tailored combination of one-hot loss, information loss, and activation loss, all customized according to classification characteristics. This data-free learning method (DAFL) introduces an innovative framework and manages to maintain an accuracy drop of less than 5% on both CIFAR-10 and CIFAR-100 datasets. In a similar vein, (Micaelli and Storkey 2019) and (Fang et al. 2019) employ generators to create training images. However, their approach differs from DAFL in that they view the generation and distillation processes as intertwined, making the generator responsible for producing images that encourage a discrepancy between the student's and teacher's outputs. Simultaneously, the student network is trained to mimic the teacher network.

In particular, data-free knowledge distillation places greater requirements on the generated data for the following reasons: the generated data must exhibit a wide range of variations to ensure that the student model can acquire comprehensive knowledge from this diverse dataset.

Contrastive Learning is an effective way to address model collapse issues where the synthesized instances are highly similar to each other and thus show limited effectiveness for downstream tasks (Fang et al. 2021). Contrastive learning has made significant advancements in the realm of self-supervised learning (He et al. 2020a). Its core idea is to treat every sample as a unique category and focus on teaching the model how to differentiate between them. In this work, we take a fresh look at the contrastive learning framework from a different angle, leveraging its capacity for instance discrimination to capture the diversity of generated data.

Method

3.1 Data-Free Learning for Super-resolution

DFSR(Zhang et al. 2021) demonstrates that utilizing a data-free knowledge distillation framework for super-resolution networks not only safeguards user privacy but also provides superior compressed models.

Training Samples Generation In super-resolution tasks, models take low-resolution images as input and output high-resolution images. Denote $\mathcal{G}$ as the generator to produce training samples, given a random variable z from a distribution p_z as input, the image synthesized by the generator network is $\mathcal{G}(z)$. The super-resolution result of $\mathcal{G}(z)$ using teacher network $\mathcal{T}$ is $\mathcal{T}(\mathcal{G}(z))$. Then we rescale $\mathcal{T}(\mathcal{G}(z))$ to the size of $\mathcal{G}(z)$ and get $R(\mathcal{T}(\mathcal{G}(z)))$. Generator $\mathcal{G}$ is expected to produce samples which follow the distribution of the dataset and for a dataset image I_L, its I_{SL} stays consistent with itself, then $R(\mathcal{T}(\mathcal{G}(z)))$ should be consistent with $\mathcal{G}(z)$. Therefore we propose a reconstruction loss for the generator, which is formulated as Eq. (1):

$$\mathcal{L}_R = E_{z \in p_z(z)} \left[\frac{1}{n} \| R(\mathcal{T}(\mathcal{G}(z))) - \mathcal{G}(z) \|_1 \right] \quad (1)$$

We use the adversarial loss to distill from teacher super-resolution networks without access to the original or related datasets. The generator network is optimized to produce hard samples to maximize the model discrepancy between teacher and student. The adversarial loss $\mathcal{L}_{GEN}$ is formulated as:

$$\mathcal{L}_{GEN} = -\log\left(\mathcal{L}_{KD} + 1\right) \quad (2)$$

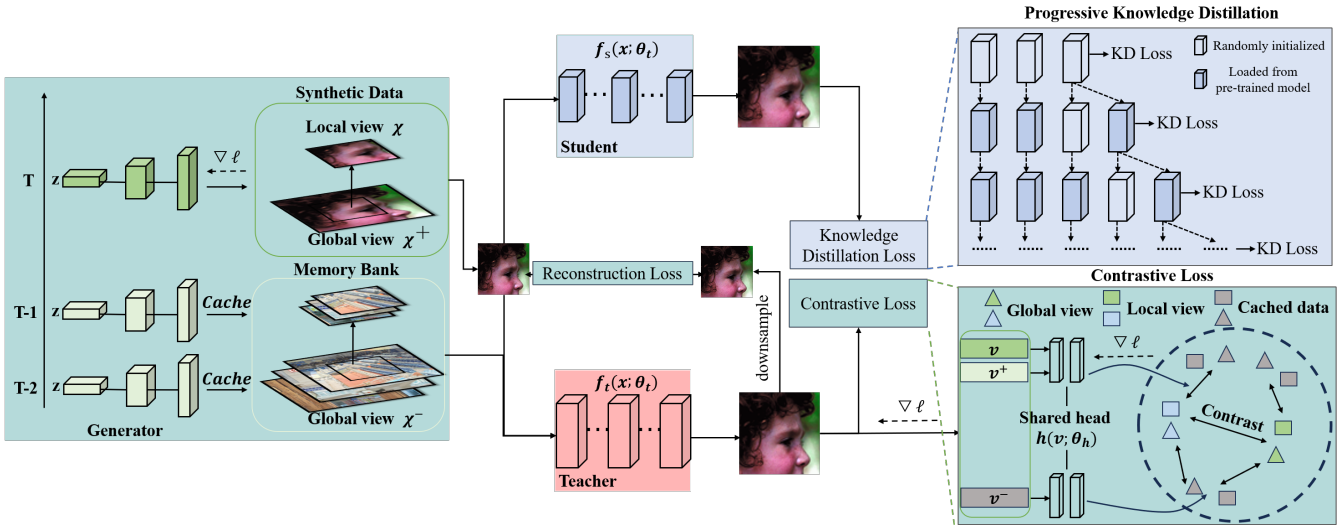

Figure 2: Framework of data-free knowledge distillation based on contrastive learning. The generator is trained with reconstruction loss, Contrastive Loss, and adversarial loss to synthesize images that are similar to the original data. The student network is then obtained utilizing progressive distillation from the teacher network.

$\mathcal{L}_{KD}$ is formulated as:

$$\mathcal{L}_{KD} = E_{x \in p_x(x)} \left[\| \mathcal{T}(x) - \mathcal{S}(x) \|_1 \right] \tag{3}$$

where x is the training sample and $p_x(x)$ is the distribution of the original dataset. Therefore, the loss function to optimize the generator can be formulated as:

$$\mathcal{L}_G = \mathcal{L}_{GEN} + \mathcal{W}_R \mathcal{L}_R \tag{4}$$

where $\mathcal{W}_R$ is the trade-off hyper-parameter to balance the two terms.

Progressive Distillation Using the progressively distill-and-refine approach involves initially training a small student network and then gradually increasing the number of layers or blocks. More specifically, for a super-resolution network $\mathcal{S}$, its function can be formulated as $\mathcal{S}(x) = \mathcal{S}_T\left(\mathcal{S}_B\left(\mathcal{S}_H(x)\right)\right)$, where x indicates the input of $\mathcal{S}$, $\mathcal{S}_H$, $\mathcal{S}_B$ and $\mathcal{S}_T$ indicate the head, body and tail of S respectively. Given that $\mathcal{S}_B$ contains N layers or blocks, we can split it into P parts $\{B_i\}_{0 \le i < P, i \in \mathbb{N}^+}$ and train $\{B_i\}_{0 \le i < P, i \in \mathbb{N}^+}$ in P steps. Initially, based on $\mathcal{S}_H$, $\mathcal{S}_T$ and B_0, we build a network $\mathcal{S}_0$ and initialize it randomly. The function of $\mathcal{S}_0$ can be formulated as $\mathcal{S}(0) = \mathcal{S}_T\left(\mathcal{B}_0\left(\mathcal{S}_H(x)\right)\right)$. In the process of training $\mathcal{S}_0$, the knowledge distillation loss is $\mathcal{S}_0$ formulated as:

$$\mathcal{L}_{KD_{S_i}} = E_{z \in p_z(z)} \left[\frac{1}{n} \left\| \mathcal{T}(G(z)) - \mathcal{S}_i(G(z)) \right\|_1 \right] \tag{5}$$

where $0 \le i < P, i \in \mathbb{N}$. After training $\mathcal{S}_0$ for several steps, we add $\mathcal{B}_1$ into $\mathcal{S}_0$ and get $\mathcal{S}_1$, which performs as $\mathcal{S}_1 = \mathcal{S}_T\left(B_1\left(B_0\left(\mathcal{S}_H(x)\right)\right)\right)$. Then when training $\mathcal{S}_1$,we initialize $\mathcal{S}_1$ with trained $\mathcal{S}_0$ and use the strategy of training $\mathcal{S}_0$ to train $\mathcal{S}_1$.

3.2 Contrastive Model Inversion

The data-free knowledge distillation method faces difficulties in ensuring the diversity of training data generated within the same training batch and across different batches. Thus, we have modified the data generation approach of the DFSR generator, introducing Contrastive Model Inversion (CMI) to enhance the differences between the samples generated in each instance and the historical samples in the memory bank, thereby increasing sample diversity.

Model inversion, as a vital step for data-free knowledge distillation, aims to recover training data $\mathcal{X}^{'}$ from a pre-trained teacher model $f_t\left(x;\theta_t\right)$ as an alternative to the inaccessible original data $\mathcal{X}$. (Choi et al. 2020) combined three inversion frameworks for data-free knowledge distillation:

$$\mathcal{L}_{inv} = \alpha \cdot \mathcal{L}_{bn}(x) + \beta \cdot \mathcal{L}_{cls}(x) + \gamma \cdot \mathcal{L}_{adv}(x) \tag{6}$$

where α, β and γ are balance terms for different criteria. $\mathcal{L}_{bn}$, $\mathcal{L}_{cls}$, $\mathcal{L}_{adv}$ refer to Eq. (7), (8) and (9) respectively.

$$\mathcal{L}_{bn}(x) = \sum_l D\left(\mathcal{N}\left(\mu_l(x), \sigma_l^2(x)\right), \mathcal{N}\left(\mu_l, \sigma_l^2\right)\right) \tag{7}$$

$$\mathcal{L}_{cls}(x) = CE\left(f_t(x), c\right) \tag{8}$$

$$\mathcal{L}_{adv}(x) = -KL\left(f_t(x)/\tau \| f_s(x)/\tau\right) \tag{9}$$

The aforementioned loss does not account for the diversity of generated samples. CMI models data diversity by addressing instance discrimination in the problem. First, given a set of data $\mathcal{X}^{'}$, an intuitive description of data diversity would be "how distinguishable are the samples from the dataset", which reveals a positive correlation between the diversity and instance distinguishability. Thus if we have a certain metric $d\left(x_1, x_2\right)$ to estimate the distinguishability for an instance pair $\{x_1, x_2\}$, then we can develop a clear definition for data diversity as the following:

$$\mathcal{L}_{div}(\mathcal{X}) = \mathbb{E}_{x_1, x_2 \in \mathcal{X}}\left[d\left(x_1, x_2\right)\right] \tag{10}$$

where $d(x_1, x_2)$ will be applied to all possible (x_1, x_2) pairs from $\mathcal{X}$.

We introduce another network $h(\cdot)$ as an instance discriminator upon the teacher network f_t that accepts feature $f_t(x)$ as input and projects it into a new embedding space. For simplification, we use $v = h(x)$to represent $v = (h \circ f_t)(x)$ because the teacher network is fixed. In the new embedding space of $h(\cdot)$, we use simple cosine similarity to describe the relationship between data pair x1 and x2 as the following:

$$sim(x_1, x_2, h) = \frac{\langle h(x_1), h(x_2) \rangle}{\|h(x_1)\| \cdot \|h(x_2)\|} \tag{11}$$

For each instance x $\in \mathcal{X}'$, we construct a positive view x^+ by random augmentation and treat other instances x^- as negative ones. The contrastive learning loss is formalized as follows:

$$\mathcal{L}_{cr}(\mathcal{X}, h) = -\mathbb{E}_{x_i \in \mathcal{X}} \left[\log \frac{\exp\left(sim\left(x_i, x_i^+, h\right)/\tau\right)}{\sum_j \exp\left(sim\left(x_i, x_j^-, h\right)/\tau\right)} \right] \tag{12}$$

where τ denotes the temperature. Discriminator $h(\cdot)$ can learn how to distinguish different samples by pushing positive pairs closer and pulling negative pairs apart, which provides a "contrast" metric for any $d(x_1, x_2)$ pair.

3.3 Optimization

By combining all the aforementioned loss functions and the progressive distillation method, we obtain the final objective functions (13) for the generator and (14) for the student respectively.

$$\mathcal{L}_G = \mathcal{L}_{GEN} + \mathcal{W}_R \mathcal{L}_R + \mathcal{W}_C \mathcal{L}_{cr} \tag{13}$$

where $\mathcal{W}_R \mathcal{W}_Cand\mathcal{W}_I$ is the hyper-parameter.

$$\mathcal{L}_S = \mathcal{L}_{KD_{S_{P-1}}} \tag{14}$$

Our training strategy is summarized as Algorithm. 1. We apply an iterative and progressive training strategy to optimize generator } and student network $\mathcal{S}$. While given student network $\mathcal{S}$ and the number of segments of student network body P , we construct a set of tiny student networks $\{S_i(x; \theta^s)\}_{0 \le i < P, i \in \mathbb{N}^+}$. Then we train S_i in turn and initialize S_{i+1} with trained S_i. The training process of S_i and $\mathcal{G}$ are performed alternately. In one iteration, we fix the generator $\mathcal{G}$, calculate knowledge distillation loss with Eq. (15), and then update the parameter of S_i via backward propagation. After updating S_i for several steps, we fix the parameter of S_i and calculate Eq. (13) to optimize $\mathcal{G}$. It's worth noting that when we start training S_{i+1}, the generator $\mathcal{G}$ will not be reinitialized. After all the networks in $\{S_i(x; \theta^s)\}_{0 \le i < P, i \in \mathbb{N}^+}$ are trained according to the preceding procedure, the training process of our student network $\mathcal{S}$ is complete.

Algorithm 1: Data-Free Knowledge Distillation For Image Super-Resolution Based On Contrastive Learning

Input: A pre-trained super-resolution teacher model $\mathcal{T}$; P indicates the number of student body segments; M denotes batch size; $p(z)$ denotes noise prior.

Initialize: Randomly initialize a student model $\mathcal{S}(x; \theta^s)$ and a generator $\mathcal{G}(x; \theta^g)$
Initialize Set $\{S_i(x; \theta^s)\}_{0 \le i < P, i \in \mathbb{N}^+}$ based on $\mathcal{S}(x; \theta^s)$ and $\mathcal{G}(x; \theta^g)$ randomly.
$\mathcal{B} \leftarrow \emptyset$
Initialize discriminator $h(\cdot; \theta_h)$
for $k = 0$ to P **do**
    Initialize $\mathcal{S}_k \leftarrow \mathcal{S}_{\max(k-1,0)}$.
    for number of training iterations **do**
        Imitation Stage:
        for k steps **do**
            Sample noise images $\{z^i\}_{i \le M}$ from $p(z)$.
            Get generated images $\{\mathcal{G}(z^i)\} \leftarrow \{z^i\}$.
            Obtain SR results $\{\mathcal{T}(\mathcal{G}(z^i))\}$, $\{\mathcal{S}_k(\mathcal{G}(z^i)\}$.
            Calculate loss $\mathcal{L}_{S_k}$ via Eq. (14).
            Update θ^{sk} with $\nabla \mathcal{L}_{S_k}$.
        end for
        Generation Stage:
        $x \leftarrow \mathcal{G}(z; \theta_g)$
        $x_B \leftarrow$ sample($\mathcal{B}$)
        Calculate loss $\mathcal{L}_{\mathcal{G}}$ with Eq. (13).
        $z \leftarrow z - \eta \nabla_z \mathcal{L}_{\mathcal{G}}$
        $\theta_g \leftarrow \theta_g - \eta \nabla_{\theta_g} \mathcal{L}_{\mathcal{G}}$
        $\theta_h \leftarrow \theta_h - \eta \nabla_{\theta_h} \mathcal{L}_{\mathcal{G}}$
    end for
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{x\}$
end for=0

Output: Output the trained student network $\mathcal{S}(x; \theta^s)$.

Experiments

In this section, we conduct extensive experiments to compare the effectiveness of the proposed data-free distillation method and our method on various super-resolution datasets. Quantitative and qualitative results are compared with baselines of VDSR.

4.1 Baselines

A bunch of baselines are compared to demonstrate the effectiveness of our proposed method. The baselines are briefly described as follows.

Teacher: the given pre-trained model which serves as the teacher in the distillation process.

DFSR: the student trained using the methodology described in DFSR.

Ours: the student trained with images generated through CMI.

4.2 Experiments on VDSR

Firstly, we experiment with our method on VDSR (Kim, Lee, and Lee 2016). We choose the VDSR model as the teacher super-resolution model and then halve the number of channels in the teacher network to get our student network (denoted as VDSR-half). we use 291 images as in for training and Set5 for validation in our experiments. The method is implemented based on the open-source Pytorch code of VDSR. Optimizers for the student and generator are SGD and Adam.

We use three generators corresponding to three super-

Table 1: Quantitative results (PSNR/SSIM) of VDSR in different experimental settings.

Dataset	Scale	VDSR							
		Teacher		Bicubic		DFSR		Ours	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM
Set5	x2	37.65	0.9046	33.69	0.8475	37.13	**0.8886**	**37.22**	0.8857
	x3	33.77	0.8221	30.40	0.7337	33.32	0.8005	**33.41**	**0.8019**
	x4	31.46	0.7404	28.41	0.6323	30.99	0.7175	**31.08**	**0.7191**
Set14	x2	33.15	0.8282	30.33	0.7623	32.80	**0.8151**	**32.84**	0.8132
	x3	29.87	0.7124	27.23	0.6317	29.55	**0.6922**	**29.58**	0.6918
	x4	30.45	0.7109	28.03	0.6311	30.16	0.6945	**30.18**	**0.6947**

GT Bicubic Teacher DFSR Ours

Figure 3: ×4 super resolution results of head from Set5 on VDSR

resolution scales to generate images for distillation. During each training update, we randomly select a scale among ×2, ×3, and ×4. Only the generator corresponding to the selected scale constructs the mini-batch and undergoes an update. The student network and all generators are randomly initialized. Throughout the optimization process, the learning rate for the student network is initially set to 0.1 and then decreased by a factor of 10 every 10 epochs. As for the generators, the initial learning rate is set to 1e-5 and decayed following the same strategy as the student network.

Table 1 shows the performance of the student model obtained with different methods. In this table, Teacher indicates the pre-trained teacher model, and DFSR indicates the original method. As is shown in the table, our method performs significantly better and achieves results close to training with the original dataset. The visual qualities of the same architecture using different training strategies are shown in Figure 2. Our method shows similar visual quality with students trained with the original dataset and performs better than training using DFSR and bicubic results.

Conclusion

In this work, we introduce Contrastive Model Inversion (CMI) to the vanilla DFSR generator to guarantee the diversity of synthetic data, which can bring significant benefits for downstream distillation tasks. Then, the reconstruction loss and the adversarial loss are utilized to train the generator for approximating the original training data as well as making a difference in the results of teachers and students. Furthermore, we adopt a Progressive Knowledge Distillation training strategy to distill additional insights from the teacher network and enhance the training of the student network. Extensive experiments demonstrate that our method can produce student networks with better results without training data, which meets the urging demand of resource-constrained devices. In addition, the method can be transferred to other tasks such as image denoising and inpainting with a similar framework. Finally, the field of Data-Free Knowledge Distillation remains to be explored, our work only glimpses the tip of this area and may it provide inspiration for future works.

References

Chen, H.; Yunhe, W.; Xu, C.; Yang, Z.; Liu, C.; Shi, B.; Chunjing, X.; Xu, C.; and Tian, Q. 2019. DAFL: Data-Free Learning of Student Networks.

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

Chowdhary, K.; and Chowdhary, K. 2020. Natural language processing. *Fundamentals of artificial intelligence*, 603–649.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11065–11074.

Fang, G.; Song, J.; Shen, C.; Wang, X.; Chen, D.; and Song, M. 2019. Data-Free Adversarial Distillation. *Cornell University - arXiv,Cornell University - arXiv*.

Fang, G.; Song, J.; Wang, X.; Shen, C.; Wang, X.; and Song, M. 2021. Contrastive Model Inversion for Data-Free Knowledge Distillation. *Cornell University - arXiv,Cornell University - arXiv*.

Gao, Q.; Zhao, Y.; Li, G.; and Tong, T. 2019. *Image Super-Resolution Using Knowledge Distillation*, 527–541.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020a. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, Z.; Dai, T.; Lu, J.; Jiang, Y.; and Xia, S.-T. 2020b. Fakd: Feature-Affinity Based Knowledge Distillation for Efficient Image Super-Resolution. In *2020 IEEE International Conference on Image Processing (ICIP)*.

Hui, Z.; Wang, X.; and Gao, X. 2018. Fast and Accurate Single Image Super-Resolution via Information Distillation Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hunt, B. R. 1995. Super-resolution of images: Algorithms, principles, performance. *International Journal of Imaging Systems and Technology*, 6(4): 297–304.

Jiang, K.; Wang, Z.; Yi, P.; Jiang, J.; Xiao, J.; and Yao, Y. 2018. Deep Distillation Recursive Network for Remote Sensing Imagery Super-Resolution. *Remote Sensing*, 1700.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1646–1654.

Micaelli, P.; and Storkey, A. 2019. Zero-shot Knowledge Transfer via Adversarial Belief Matching. *Cornell University - arXiv,Cornell University - arXiv*.

Nayak, G.; Mopuri, K.; Shaj, V.; Babu, R.; and Chakraborty, A. 2019. Zero-Shot Knowledge Distillation in Deep Networks. *Cornell University - arXiv,Cornell University - arXiv*.

Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, K.; Cao, X.; and Shen, H. 2020. Single image super-resolution via a holistic attention network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 191–207. Springer.

Pak, M.; and Kim, S. 2017. A review of deep learning in image recognition. In *2017 4th international conference on computer applications and information processing technology (CAIPT)*, 1–3. IEEE.

Wang, X.; Yi, J.; Guo, J.; Song, Y.; Lyu, J.; Xu, J.; Yan, W.; Zhao, J.; Cai, Q.; and Min, H. 2022. A review of image super-resolution approaches based on deep learning and applications in remote sensing. *Remote Sensing*, 14(21): 5423.

Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.-H.; and Liao, Q. 2019. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12): 3106–3121.

Zhang, Y.; Chen, H.; Chen, X.; Deng, Y.; Xu, C.; and Wang, Y. 2021. Data-free knowledge distillation for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7852–7861.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11953–11962.