FAKD: FEATURE-AFFINITY BASED KNOWLEDGE DISTILLATION FOR EFFICIENT IMAGE SUPER-RESOLUTION

Zibin He[⋆], *Tao Dai*[⋆,†], *Jian Lu*[‡], *Yong Jiang*[⋆,†], *Shu-Tao Xia*[⋆,†]

[⋆]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[†] PCL Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen, China
[‡] Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Shenzhen University, China
hzb19@mails.tsinghua.edu.cn, daitao.edu@gmail.com, {jiangy,xiast}@sz.tsinghua.edu.cn

ABSTRACT

Convolutional neural networks (CNNs) have been widely used in image super-resolution (SR). Most existing CNN-based methods focus on achieving better performance by designing deeper/wider networks, while suffering from heavy computational cost problem,thus hindering the deployment of such models in mobile devices with limited resources. To relieve such problem, we propose a novel and efficient SR model, named Feature Affinity-based Knowledge Distillation (FAKD), by transferring the structural knowledge of a heavy teacher model to a lightweight student model. To transfer the structural knowledge effectively, FAKD aims to distill the second-order statistical information from feature maps and trains a lightweight student network with low computational and memory cost. Experimental results demonstrate the efficacy of our method and the effectiveness over other knowledge distillation based methods in terms of both quantitative and visual metrics.

***Index Terms*—** Image super-resolution, Knowledge distillation, Model compression, Convolutional neural networks

1. INTRODUCTION

Single image super-resolution (SISR) [1] aims to generate a high-resolution (HR) image from its degraded low-resolution (LR) counterpart. It can be utilized in a variety of computer vision applications, such as object recognition [2], medical imaging [3] and image generation [4]. To date, a plenty of SISR methods have been developed, including interpolating based [5], sparse representation based [6] and deep learning based methods [7, 8].

Nowadays, convolutional neural networks (CNNs) have prevailed in image super-resolution task and achieved impressive performance due to the powerful ability of feature representational expression. In a pioneer work in [9], Dong et al. first proposed an end-to-end convolution neural network (SRCNN) to learn the mapping function between LR images and their corresponding HR images. Later works like EDSR [10], RDN[11] and RCAN [12] build very deep networks by stacking residual blocks to hundreds of layers to achieve the state-of-the-art results. However, most existing CNN-based SR methods suffer from heavy computational cost problem, as they contain a huge number of parameters. In practice, such heavy models are limited in real applications due to the difficulty of the deployment in resource-limited devices, such as mobile phones and robots. Thus, it is crucial to design lightweight SR models.

To obtain lightweight models, there exist recent attempts to compress network models, including model pruning [13, 14, 15], lightweight network design [16, 17] and knowledge distillation (KD) methods [18, 19, 20, 21]. Model pruning and lightweight network design methods require elaborate design and may result in performance degradation. By contrast, knowledge distillation methods contain advantages over other model compression methods without changing the network structure.

Traditional knowledge distillation (KD) [18] is first proposed for image recognition tasks and follows a teacher-student paradigm by using soft labels of a strong teacher network to supervise the training of a tiny student network. So far, several KD methods have been proposed. For example, Romero et al. proposed FitNet [19] to distill the knowledge hidden in the feature maps of intermediate layers. Sergey et al. [22] proposed attention transfer by calculating attention maps from mid-level features. The student network is encouraged to generate similar attention maps as teachers. Considering the significance of the correlation between layers, Yim et al. [20] proposed Flow of Solution Procedure (FSP) to extract the problem-solving information and regard it as a supervisory signal for training the student. However, most existing KD methods focus on high-level tasks, like image classification [19, 22, 20], while little attention has been devoted to image regression task, like image SR. When confronting image SR, it is still an open problem of how to compress models, since the representation space is unbounded [23, 24]. Here, we attempt to design a KD based framework for efficient SR.

For efficient SR, we propose a novel feature affinity-based knowledge distillation (FAKD) framework by distilling the structural knowledge from a teacher model. The most related work [21] attempts to propagate the simple first-order statistical information (e.g., average pooling over channels) from a teacher model, while neglecting the rich high-order statistical information. For this reason, we focus on distilling the second-order information (e.g., intra-feature correlation) from feature maps, which is shown to be helpful for more accurate reconstruction [7]. Specifically, FAKD transfers the knowledge from feature correlation map of a teacher model to a lightweight student model, which forces the lightweight student model to mimic the feature correlation. Experiments demonstrate that our proposed framework effectively compresses the CNN-based SR models, while improving the performance of student network by transferring the structural knowledge from a strong teacher model.

Corresponding author: Tao Dai

This work is supported in part by the National Natural Science Foundation of China under Grant 61771273, Guangdong Basic and Applied Basic Research Foundation on 2019A1515110344, the China Postdoctoral Science Foundation under Grant 2019M660645, the RD Program of Shenzhen under Grant JCYJ20180508152204044, and the project "PCL Future Greater-Bay Area Network Facilities for Large-scale Experiments and Applications (LZC0019)".

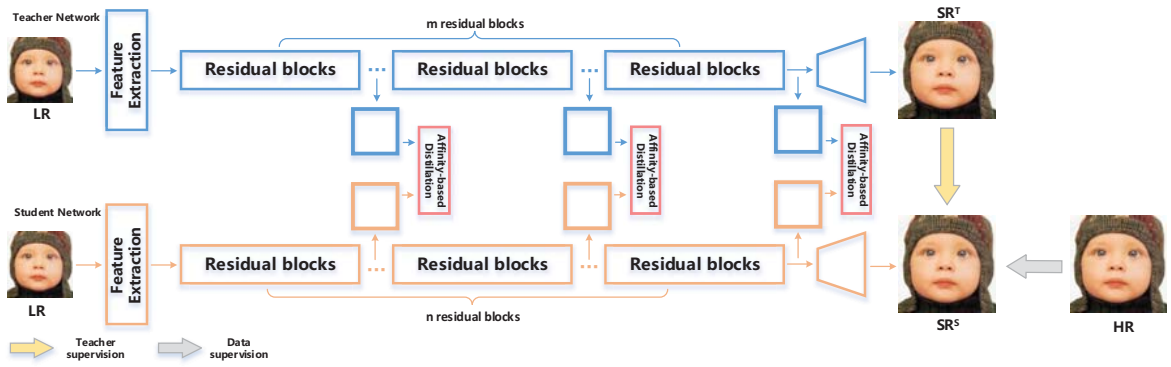

Fig. 1. The architecture of our feature affinity-based knowledge distillation (FAKD) framework for efficient image super-resolution. Given a heavy teacher and its corresponding lightweight student network, FAKD transfers the structural knowledge from the strong teacher model to the lightweight student model by forcing similar feature correlation maps between the teacher and student models.

In summary, the main contributions are summarized as follows:

- We propose a feature affinity-based knowledge distillation (FAKD) framework, which leverages the correlation within a feature map to supervise the training of student networks. Affinity information in spatial dimension is explored to improve the distillation performance.
- Experiments revealed the superiority of our proposed framework in terms of both quantitative and visual results.

2. PROPOSED METHOD

The pipeline of our proposed feature affinity-based knowledge distillation framework is shown in Figure 1. The degraded LR images propagate through both teacher T and student S network. Teacher model is a powerful cumbersome network while student model is a lightweight network. In our framework, both of them share the same architecture with different hyperparameter (e.g., network depth). As shown in Figure 1, they are composed of m and n residual blocks ($m > n$), respectively. In order to transfer the knowledge from the teacher model to the student model effectively, the intermediate feature maps of student network are forced to mimic the feature affinity matrix from the teacher model. Furthermore, teacher output images and ground-true images are also used to supervise student network via teacher supervision (TS) and data supervision (DS) respectively.

2.1. Feature Affinity-based Distillation (FAKD)

The key to knowledge distillation is to design an appropriate mimicry loss function that can successfully propagate valuable information to guide the training process of the student model. Previous research [23, 24] have demonstrated that the feature representation space of regression problems is unbounded. For this reason, the existing distillation approaches [19, 22, 20] designed for classification tasks may not be suitable for image SR due to the huge solution space. To render knowledge distillation effectively for image SR, it is necessary to limit the solution space. To this end, we design a general feature affinity-based knowledge distillation framework for efficient SR.

Given a batch of feature maps $F \in R^{b\times C\times W\times H}$, we firstly reshape them into a three-dimensional tensor $F \in R^{b\times C\times WH}$, which are instance, channel and spatial dimension, respectively. In order to exploit congruity within feature maps, we propose to calculate affinity matrix A. They are generated using feature maps from low-level, mid-level and high-level layers to represent different levels of correlation. The student network is encouraged to produce similar affinity matrices with teacher networks and the feature affinity-based distillation loss can be formulated as

$$L_{AD} = \frac{1}{|A|}\sum_{l=1}^{l'} ||A_l^S - A_l^T||_1, \tag{1}$$

where A_l^T and A_l^S are the affinity matrix of teacher and student network extracted from the feature maps of the l-th layer; l' is the number of layer we choose to extract. $|A|$ denotes the number of elements in the affinity matrix.

To preserve the spatial contiguity among pixels, we consider the affinity matrix from spatial perspective, aiming to explore the relationship between pixels. The pipeline is illustrated in Figure 2, where every pixel is regarded as a C-dimensional vector, the blue column, and normalization is conducted across every column, seen in Equation 2. After normalization, every column is unit-length, therefore the cosine similarity between two pixels is simply obtained by inner product, which empirically works well. The spatial affinity matrix is formulated as:

$$\tilde{F}_{[i,:,j]} = \frac{F_{[i,:,j]}}{||F_{[i,:,j]}||_2^2}, \tag{2}$$

$$A = \tilde{F}^T \cdot \tilde{F}, \tag{3}$$

where $\tilde{F}$ is the normalized feature maps. The size of generated spatial affinity matrix is $b \times HW \times HW$. Every element in spatial affinity matrix represents the spatial correlation between two pixels.

2.2. Overall Loss Function

Along with feature affinity-based distillation, we empirically found that teacher supervision (TS) and data supervision (DS) also help improve the distillation performance, as shown in Figure 1. TS and DS are supposed to compare student's output with teacher's and ground-true images respectively, seen in Equation 4 and 5. Therefore, student network can receive supervisory signals from both teacher distribution and real-data distribution. The overall loss function is formulated as Equation 6.

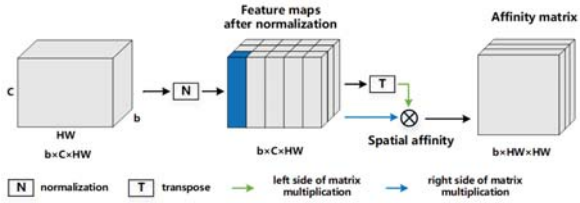

Fig. 2. Spatial affinity calculation pipeline with C the number of channel, H, W the spatial size, b the batch size. The blue pixels are the dimension of affinity, and will be normalized.

$$L_{TS} = \sum_{i=1}^{N} ||I_{SR}^{S(i)} - I_{SR}^{T(i)}||_1, \tag{4}$$

$$L_{DS} = \sum_{i=1}^{N} ||I_{SR}^{S(i)} - I_{HR}^{(i)}||_1, \tag{5}$$

$$L(\theta) = \alpha L_{DS} + \beta L_{TS} + \gamma L_{AD}, \tag{6}$$

where I_{SR}^{S}, I_{SR}^{T} and I_{HR} are the images from student outputs, teacher outputs and ground-true set respectively. α, β and γ are penalty coefficients to balance different aspects of loss. Using this overall loss function, the student network can be optimized to capture all these knowledge from the teacher.

3. EXPERIMENTAL RESULTS

3.1. Experimental Settings

Following [10, 12], 800 images from DIV2K [25] are used as training set. For testing, we used four benchmark datasets: Set5 [26], Set14 [27], BSD100 [28], Urban100 [29]. LR images are obtained by Bicubic interpolation (BI) [30], and PSNR and SSIM [31] serve as the evaluation metrics.

We use RCAN [12] and SAN [7] as our main teacher and student networks to verify the effectiveness of our general distillation framework. The network configurations are demonstrated in Table 1. RCAN is composed of multiple residual groups (resgroup), each of which contains numerous residual block (resblock) inside. We decrease the number of residual block so that the amount of total parameters in student network drops to approximately 30% of teacher. SAN is also built by stacking residual groups and blocks, and it modifies the attention module of RCAN into a second-order version. Like RCAN, We also decrease the number of residual group from 20 to 6.

The student model is trained with ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$. The initial learning rate is set to 10^{-4} and decreased to half every 150 epochs.

Table 1. Network configurations of teacher and student networks. $\#A$ denotes the number of A. TN and SN are teacher and student network respectively.

Network		#resgroup	#resblock	#channel
RCAN	TN	10	20	64
	SN	10	6	64
SAN	TN	20	10	64
	SN	6	10	64

Table 2. The effects of different components in the loss function. DS = data supervision, TS = teacher supervision, SA = spatial affinity. The best result is highlighted in bold type.

Component			PSNR			
DS	TS	SA	Set5	Set14	BSD100	Urban100
✓			32.321	28.709	27.634	26.340
✓	✓		32.362	28.722	27.657	26.382
✓		✓	32.410	28.731	27.671	26.394
✓	✓	✓	**32.462**	**28.750**	**27.678**	**26.422**

3.2. Ablation Study

The effectiveness of distillations. To verify the effect of different components in the loss function, we perform RCAN with various loss functions. The overall results are shown in Table 2. In the 1^{st} row, only data supervision is adopted, which means that the student is trained without distillation, using only HR images as supervision. Student network in other three rows are all trained using different distillation strategies. When teacher supervision (TS) or spatial affinity (SA) is added respectively(2^{nd} and 3^{rd} row), the performance can be improved. We further investigate the effect of the overall distillation framework. As shown in the last row in Table 2, our method obtains the best results. Comparing with the base model without knowledge distillation (the 1^{st} row), knowledge distillation strategy can achieve consistent performance gain on various datasets. In the following experiments, we use the combination of DS, TS and SA as our default experimental setting.

Comparison with other feature KD methods. We replace our feature affinity-based distillation with five variants to verify the superiority of our distillation scheme. The variants include FitNet, Attention Transfer(AT) and Flow of Solution Procedure(FSP), channel affinity(CA) and instance affinity(IA).

- Feature distillation by FitNet [19]: We follow [19] to directly align feature maps between teacher and student network.
- Feature distillation by AT [22]: We aggregate feature maps in channel dimension to generate attention maps, then transfer them from teacher to student network.
- Feature distillation by FSP [20]: Gram matrix calculated between two intermediate layers is distilled.
- Feature distillation by channel affinity (CA): The mechanism is similar to spatial affinity, and we view a channel as a HW-dimension vector. The normalization is conducted in every channel and affinity between two channels is obtained by employing similar method as Equation 3. The resulted channel affinity matrix is a $b \times C \times C$ matrix.
- Feature distillation by instance affinity(IA): An instance is regarded as a CWH-dimensional vector. Similarly, Normalization is followed in every instance and correlation between instances can be acquired. The size of instance affinity matrix is $b \times b$.

From Table 4, we can see that our spatial affinity method outperforms all the other distillation strategies. The superiority over FitNet can be attributed to the bounded representation space after our transformation. The superiority over AT and FSP comes from that we translate the knowledge from teacher network into a compressed space that is more informative, with the capability of capturing affinity information from spatial perspective. Comparing with other two affinity-based distillation schemes, spatial affinity is much effective because the information extracted in spatial domain is substantially richer than that in channel and instance domain.

Table 3. Quantitative results (PSNR/SSIM) in different experimental settings. The best results are highlighted.

Datasets	Scale	RCAN			SAN		
		TN	SN w/o FAKD	SN with FAKD	TN	SN w/o FAKD	SN with FAKD
Set5	×2	38.271/ 0.9614	38.074/ 0.9608	**38.164/ 0.9611**	38.310/ 0.9620	38.059/ 0.9607	**38.168/ 0.9611**
	×3	34.758/ 0.9299	34.557/ 0.9284	**34.653/ 0.9291**	34.750/ 0.9300	34.527/ 0.9278	**34.644/ 0.9291**
	×4	32.638/ 0.9002	32.321/ 0.8964	**32.462/ 0.8982**	32.640/ 0.9003	32.360/ 0.8968	**32.472/ 0.8980**
	×8	27.310/ 0.7878	27.013/ 0.7758	**27.154/ 0.7800**	27.170/ 0.7829	26.998/ 0.7761	**27.121/ 0.7797**
Set14	×2	34.126/ 0.9216	33.623/ 0.9183	**33.815 /0.9190**	34.070/ 0.9213	33.612/ 0.9175	**33.723/ 0.9185**
	×3	30.627/ 0.8476	30.408/ 0.8438	**30.449/ 0.8442**	30.590/ 0.8476	30.385/ 0.8430	**30.482/ 0.8454**
	×4	28.851/ 0.7885	28.688/ 0.7840	**28.750/ 0.7859**	28.920/ 0.7888	28.673/ 0.7841	**28.762/ 0.7858**
	×8	25.261/ 0.6516	24.970/ 0.6402	**25.085/ 0.6441**	25.140/ 0.6476	24.996/ 0.6422	**25.064/ 0.6431**
BSD100	×2	32.390/ 0.9027	32.199/ 0.9000	**32.274/ 0.9010**	32.420/ 0.9028	32.207/ 0.9001	**32.287/ 0.9013**
	×3	29.309/ 0.8113	29.162/ 0.8076	**29.208/ 0.8087**	29.330/ 0.8112	29.140/ 0.8066	**29.213/ 0.8090**
	×4	27.748/ 0.7433	27.634/ 0.7381	**27.678/ 0.7402**	27.780/ 0.7436	27.637/ 0.7385	**27.685/ 0.7400**
	×8	24.975/ 0.6056	24.805/ 0.5971	**24.868/ 0.6001**	24.880/ 0.6011	24.802/ 0.5981	**24.859/ 0.5986**
Urban100	×2	33.176/ 0.9384	32.317/ 0.9302	**32.533/ 0.9320**	33.100/ 0.9370	32.331/ 0.9304	**32.463/ 0.9317**
	×3	29.014/ 0.8687	28.482/ 0.8600	**28.523/ 0.8602**	28.930/ 0.8671	28.421/ 0.8582	**28.540/ 0.8607**
	×4	26.748/ 0.8067	26.340/ 0.7933	**26.422/ 0.7973**	26.790/ 0.8068	26.345/ 0.7939	**26.414/ 0.7965**
	×8	22.978/ 0.6441	22.481/ 0.6175	**22.568/ 0.6226**	22.700/ 0.6314	22.494/ 0.6208	**22.629/ 0.6231**
Flops(G)		36.80	12.93	12.93	37.55	12.52	12.52
Params		15.59M	5.17M	5.17M	15.86M	5.0M	5.0M

Table 4. Comparison with other feature distillation methods.

Method	Set5	Set14	BSD100	Urban100
FitNet[19]	32.425	28.717	27.666	26.395
AT[22]	32.412	28.731	27.657	26.363
FSP[20]	32.374	28.724	27.666	26.418
CA	32.320	28.662	27.618	26.234
IA	32.403	28.713	27.646	26.386
SA	**32.462**	**28.750**	**27.678**	**26.422**

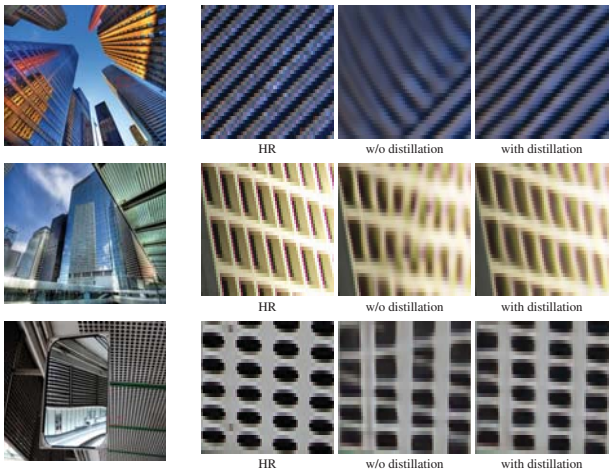

Fig. 3. Visual comparison between HR and SR with/without distillation. The best results are highlighted.

3.3. Benchmark Results

Quantitative results. The quantitative evaluation results of PSNR and SSIM in two teacher-student settings are shown in Table 3. We train our student networks in 200 epochs with batch size 16. For all datasets in four scales $\times 2$, $\times 3$, $\times 4$ and $\times 8$, our affinity-based knowledge distillation (FAKD) achieves the best performance, which indicates the efficacy and superiority of our method. The average performance gain of RCAN and SAN is approximately 0.1dB, and SAN's PSNR gain is greater than that of RCAN, which is mainly because teacher model of SAN is more powerful and thus can provide more structural knowledge. Therefore, more useful supervisory signal can be extracted from teacher SAN, leading to better performance. Model size analysis and FLOPs are also conducted. The FLOPs is measured in scale $\times 4$ with input image size of 48×48. The results reveal a substantial model size decline while performances are reasonably acceptable. Our distillation framework can steadily boost the performance while introducing no extra parameters and reducing computation amount with a substantial margin.

Visual results. In Figure 3, we also show the visual results of output images with/without knowledge distillation. From figure 3 we can see that our method with knowledge distillation scheme produces sharper results and restores more image details (e.g., the lines), while failing to recover such image details without knowledge distillation scheme. This demonstrates that our knowledge distillation scheme can effectively transfer the knowledge from the teacher model. Additionally, With our distillation scheme, blurred artifacts can be mitigated to some extent.

4. CONCLUSION

In this paper, we propose a general feature affinity-based knowledge distillation (FAKD) framework for efficient image super-resolution. In our FAKD, we take into consideration the spatial affinity among pixels within a feature map. In order to effectively transfer the rich feature knowledge from teacher model, we extract affinity knowledge from different layers to represent the scale of features. Extensive experiments demonstrate the effectiveness of our proposed method.

5. REFERENCES

[1] William T Freeman, Egon C Pasztor, and Owen T Carmichael, "Learning low-level vision," *IJCV*, 2000.

[2] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *ICCV*, 2017.

[3] Wenzhe Shi, Jose Caballero, Christian Ledig, Xiahai Zhuang, Wenjia Bai, Kanwal Bhatia, Antonio M Simoes Monteiro de Marvao, Tim Dawes, Declan O'Regan, and Daniel Rueckert, "Cardiac image super-resolution with global correspondence using multi-atlas patchmatch," in *MICCA*, 2013.

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[5] Xiangjun Zhang and Xiaolin Wu, "Image interpolation by adaptive 2-d autoregressive modeling and soft-decision estimation," *TIP*, 2008.

[6] Weisheng Dong, Lei Zhang, Rastislav Lukac, and Guangming Shi, "Sparse representation based image interpolation with nonlocal autoregressive modeling," *TIP*, 2013.

[7] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang, "Second-order attention network for single image super-resolution," in *CVPR*, 2019.

[8] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao, "Deep learning for single image super-resolution: A brief review," *TMM*, 2019.

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *TPAMI*, 2016.

[10] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPRW*, 2017.

[11] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *CVPR*, 2018.

[12] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.

[13] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.

[14] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang, "Learning efficient convolutional networks through network slimming," in *ICCV*, 2017.

[15] Yihui He, Xiangyu Zhang, and Jian Sun, "Channel pruning for accelerating very deep neural networks," in *ICCV*, 2017.

[16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[17] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *CVPR*, 2018.

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

[19] 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.

[20] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *CVPR*, 2017.

[21] Qinquan Gao, Yan Zhao, Gen Li, and Tong Tong, "Image super-resolution using knowledge distillation," in *ACCV*, 2018.

[22] 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.

[23] Feng Zhang, Xiatian Zhu, and Mao Ye, "Fast human pose estimation," in *CVPR*, 2019.

[24] Muhamad Risqi U Saputra, Pedro PB de Gusmao, Yasin Almalioglu, Andrew Markham, and Niki Trigoni, "Distilling knowledge from a deep pose regressor network," *arXiv preprint arXiv:1908.00858*, 2019.

[25] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *CVPRW*, 2017.

[26] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[27] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*, 2010.

[28] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, 2001.

[29] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, "Single image super-resolution from transformed self-exemplars," in *CVPR*, 2015.

[30] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *CVPR*, 2018.

[31] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, 2004.