MTKD: Multi-Teacher Knowledge Distillation for Image Super-Resolution

Yuxuan Jiang, Chen Feng, Fan Zhang, and David Bull

Visual Information Laboratory, University of Bristol, Bristol, BS1 5DD, UK
{yuxuan.jiang, chen.feng, fan.zhang, dave.bull}@bristol.ac.uk

Abstract. Knowledge distillation (KD) has emerged as a promising technique in deep learning, typically employed to enhance a compact student network through learning from their high-performance but more complex teacher variant. When applied in the context of image super-resolution, most KD approaches are modified versions of methods developed for other computer vision tasks, which are based on training strategies with a single teacher and simple loss functions. In this paper, we propose a novel Multi-Teacher Knowledge Distillation (MTKD) framework specifically for image super-resolution. It exploits the advantages of multiple teachers by combining and enhancing the outputs of these teacher models, which then guides the learning process of the compact student network. To achieve more effective learning performance, we have also developed a new wavelet-based loss function for MTKD, which can better optimize the training process by observing differences in both the spatial and frequency domains. We fully evaluate the effectiveness of the proposed method by comparing it to five commonly used KD methods for image super-resolution based on three popular network architectures. The results show that the proposed MTKD method achieves evident improvements in super-resolution performance, up to 0.46 dB (based on PSNR), over state-of-the-art KD approaches across different network structures. The source code of MTKD will be made available `https://github.com/YuxuanJJ/MTKD` for public evaluation.

Keywords: Image Super-Resolution · Knowledge Distillation · Multi-teacher · MTKD

1 Introduction

Image super-resolution (ISR) is an important research topic in image processing; its purpose is to create a high-resolution (HR) image with improved perceptual quality and richer spatial detail from a corresponding low-resolution (LR) version. It is widely used in applications including medical imaging [64], image restoration [39], enhancement [58], and picture coding [1]. The past decade has seen impressive performance improvements due to extensive research in this area, in particular associated with advances in deep learning techniques. Learning-based ISR approaches can be classified according to their basic network structure [2]. One major class is based on convolutional neural networks (CNNs), with

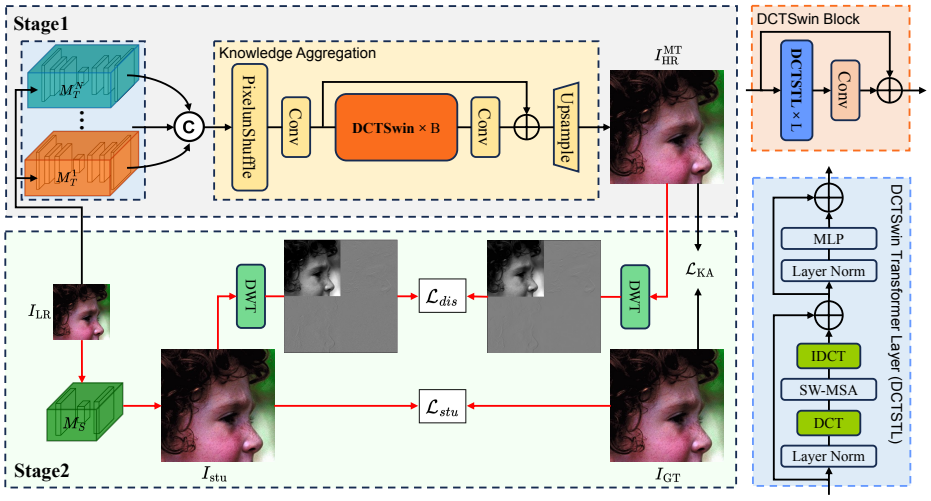

Fig. 1: Illustration of the proposed Multi-Teacher Knowledge Distillation framework.

notable examples including *e.g.*, SRCNN [14], EDSR [40] and RCAN [42, 77]. A second class employs Vision Transformer (ViT) networks with important contributions such as SwinIR [39], Swin2SR [13] and HAT [11].

Although these learning-based ISR algorithms have demonstrated superior performance over conventional methods based on classic signal processing theory, they are typically associated with high computational complexity and memory requirements, often inhibiting their practical deployment. To address this issue, research has focused on the development of lightweight ISR methods for real-world applications [6, 28]. These works often obtain compact models through model compression [4] or other simplification [39] of their corresponding full versions. However, when these lite ISR models are directly optimized based on the same training material used for their original counterparts, they generally achieve lower super-resolution performance due to their reduced model capacity and limited learning ability.

To further improve the performance of these low complexity learning-based ISR methods, knowledge distillation (KD) techniques [23] have been commonly applied. These employ a larger network as a "teacher" that transfers knowledge to a smaller "student" network. In this process, an auxiliary loss function is often employed to instruct the student to mimic the teacher's output, through minimizing the disparity between intermediate features or by aligning their final predictions. This type of approach has demonstrated promising outcomes for various applications such as image classification [30, 51], video compression [54], object detection [56, 80], natural language processing (NLP) [29, 31, 43] and quality assessment [16].

Many prior studies [15, 21, 52, 78] in KD predominantly focus on the transfer of knowledge from a single teacher to its corresponding student. This strategy loses

the advantages of well-performing complex models. They also employ simple loss objectives, such as L1, to minimize the output difference between the teacher and student models; this does not fully reflect the nature of image super-resolution in reconstructing lost high-frequency information in the higher-resolution content.

In this context, inspired by the advances in multi-teacher selection based KD for natural language processing tasks [72] and the classic wavelet transforms [22], this paper presents a novel Multi-Teacher Knowledge Distillation (MTKD) framework for image super-resolution based on a new wavelet-inspired loss function. As illustrated in Fig. 1, MTKD employs a new Discrete Cosine Transform Swin transformer (DCTSwin) based network to combine the outputs of multiple ISR teacher models and generates an enhanced representation of the high-resolution image, which is then used to guide the student model during distillation. For improving knowledge transfer performance, we have designed a new distillation loss function based on the discrete wavelet transform, which compares the output of the student and the enhanced representation within different frequency subbands. The primary contributions are summarized as follows.

1) We, for the first time, propose a novel Multi-Teacher Knowledge Distillation (MTKD) framework for image super-resolution. This framework allows the use of multiple teacher models with different network architectures, which improves the efficiency and diversity of transferred knowledge.
2) We have developed a knowledge aggregation network based on novel DCTSwin blocks, which is employed to combine the outputs of multiple teachers to produce a refined representation for ISR knowledge distillation.
3) We have designed an ISR-specific and wavelet-based loss function which collects information from different frequency subbands allowing the model to effectively learn high-frequency information from the original images, which can enhance the performance of knowledge distillation for ISR.

We have demonstrated the superior performance of the proposed method through quantitative and qualitative evaluations based on three teacher and student models with distinct network architectures. Our MTKD approach has been compared against five existing KD methods and achieved consistent and evident performance gains, up to 0.46dB assessed by PSNR.

2 Related Work

Our work is closely related to two main research topics in the literature: image super-resolution and knowledge distillation.

Image Super-Resolution (ISR) ISR is an image restoration technique, which aims to reconstruct high spatial resolution images from their low-resolution counterparts. In conventional methods, this is achieved using upsampling filters [5]. However, learning-based ISR methods are predominate due to their superior performance. One of the earliest contributions in this area is SRCNN [14], which is

based on a simple three-layer convolutional network. This has been further enhanced by incorporating more sophisticated network structures, with notable examples such as VDSR [32] based on residual connections, SRRetNet [34] and EDSR [40] employing residual blocks, and RCAN [77] containing channel weight attention mechanisms. Recently, more effective ISR algorithms have been proposed inspired by the Vision Transformer (ViT) networks [62], which leverage self-attention mechanisms for capturing extensive contextual interaction information. Important contributions include the ESRT [47], SwinIR [39] and Swin2SR [13] based on shifted window attention [44], and HAT [11] which combines both channel attention and window-based self-attention schemes. To improve the perceptual quality of super-resolution results, generative models such as variational autoencoders (VAEs), generative adversarial networks, and diffusion models have also been exploited, with typical examples including SR-VAE [45], VDVAE-SR [12], SRGAN [34], CAL-GAN [53], SR3 [55] and IDM [18]. For a more comprehensive overview of image super-resolution, readers are referred to references including [37, 69].

Knowledge Distillation (KD) The objective of Knowledge Distillation (KD) is to improve the model generalization of a compact student model by emulating the behavior of a large well-performing teacher network. It is typically applied together with model compression in order to reduce the complexity of a deep neural network [4]. KD offers enhanced performance compared to directly optimizing the compact model on training data only [23]. Existing KD approaches can be categorized into two primary classes that either: (i) perform direct emulation of the output from the teacher model (output-level) [30, 78] or (ii) allow the student network to learn intermediate features generated by the teacher model (feature-level) [15, 21, 71, 73]. These approaches have been widely applied to many high-level tasks, including, image classification [10, 30, 41, 51], object detection [7, 9, 56, 80] and natural language processing (NLP) [29, 31, 43]. In the context of ISR, SRKD is a pioneer work [17] employing KD for image super-resolution. It has been further enhanced through distilling second-order statistical information from feature maps as in FAKD [21], which improves the effective transfer of structural knowledge from the teacher to the student. In [35], the teacher model is replaced by the privileged information obtained from the ground truth high-resolution images, which has been reported to offer improved knowledge distillation performance. Moreover, knowledge distillation has been improved in [78] through image zooming and invertible data augmentations, which enhances the generalization of the student model.

It is noted that all the aforementioned knowledge distillation approaches for ISR are based on a single-teacher framework. In the context of multi-task learning, some recent studies have identified the potential of using multiple teachers for knowledge distillation [20, 25, 38, 48, 50, 70]. Similar approaches have also been proposed for applications with single tasks, such as [8, 66, 79]. These approaches involve leveraging the weighted average of teacher models for distilling into the student model with fixed weights. In contrast, [72] proposed an approach that

dynamically assigns weights to teacher models based on individual examples. However, these knowledge distillation paradigms have only been investigated for high-level tasks such as classification and detection, and have not been exploited for image super-resolution in the literature.

Moreover, existing knowledge distillation methods typically employ simple loss functions to perform observation in the spatial or feature domain. Due to the nature of the ISR task, which aims to recover spatial details (high-frequency information) in high-resolution images, it is more important to conduct an assessment in the frequency domain during the training process. It is noted that in the literature several token-mixing models have been proposed recently, which replace self-attention modules in transformers with Fourier [36] and wavelet transforms [27]. These models show evident performance enhancement for high-level vision tasks such as image classification and semantic segmentation, while maintaining low computational complexity. In the field of ISR, feature extraction has been performed in the frequency domain in order to achieve competitive performance [26,67,75], and there are also several attempts to design transform-based (DCT or Fast Fourier Transform) losses [46,68]. However, as we are aware, wavelet-based loss objectives have not previously been investigated for ISR.

3 Proposed method: MTKD

As illustrated in Fig. 1, our Multi-Teacher Knowledge Distillation (MTKD) framework consists of two primary stages: 1) knowledge aggregation and 2) model distillation. In Stage 1, the input low-resolution image $I_{\text{LR}} \in \mathbb{R}^{H \times W \times C_{\text{in}}}$ (H, W and C_{in} represent the image height, width and the number of color channels respectively) is reconstructed by N pre-trained ISR teacher models, denoted by $M_T^1 \dots M_T^N$, resulting in N high-resolution images, $I_{\text{HR}}^1 \dots I_{\text{HR}}^N$. These are then concatenated and fed into a Knowledge Aggregation module, which outputs an enhanced high-resolution image, $I_{\text{HR}}^{\text{MT}}$. Once the Knowledge Aggregation module is optimized, its parameters are then fixed in Stage 2 when the student model is trained through knowledge distillation.

3.1 Stage 1: Knowledge Aggregation

Network architecture The architecture of the network employed for knowledge aggregation is shown in Fig. 1. The input of the network is the concatenation of N teachers' outputs, $[I_{\text{HR}}^1 \dots I_{\text{HR}}^N] \in \mathbb{R}^{sH \times sW \times C_{\text{in}} \times N}$, in which s is the up-scale factor in ISR. A pixel unshuffle layer (together with an additional 3×3 convolution layer) is then used to down-sample the spatial resolution of the input by a factor of s. The down-sampling factor here is the same as the super-resolution up-scale rate in order to keep the following DCTSwin blocks independent from the upscale factor. The feature set output by the convolutional layer is denoted by $F_s \in \mathbb{R}^{H \times W \times C}$, where C represents the feature channel number that is a fixed hyperparameter. F_s are then processed by B DCTSwin blocks, generating

deep features $F_d^1, F_d^2 \dots F_d^B$. The last set of features F_d^B is fed into a 3×3 convolutional layer to produce F_d, which will be combined with the shallow features F_s before being upsampled to the full image resolution. The upsample module is implemented based on the sub-pixel convolution layer [57].

DCTSwin block As shown in Fig. 1, each DCTSwin block contains L DCTSwin Transformer Layers (DCTSTLs), which are modified from the original Swin Transformer layer [44]. It first normalizes the input using a Layer Norm before performing Discrete Cosine Transform (DCT). The transformed coefficients are then processed by the Shifted Window Multi-head Self-Attention (SW-MSA) module [44]. The output of the SW-MSA module is further fed into the Inverse DCT (IDCT) module to recover the features in the spatial domain, which are combined with the input of this DCTSTL to achieve residual learning. Here to reduce the number of model parameters and facilitate fast DCT operation, the input of the DCT is segmented into $W_s \times W_s$ blocks prior to the SW-MSA module through window partition [44], and the window reverse operation is performed after the IDCT module to reshape and assemble the output.

Loss function To train the Knowledge Aggregation module, we optimize its network parameters by minimizing the L1 loss between the ground-truth image, I_{GT}, and the output of the module, $I_{\mathrm{HR}}^{\mathrm{MT}}$:

$$\mathcal{L}_{\mathrm{KA}} = \left\| I_{\mathrm{GT}} - I_{\mathrm{HR}}^{\mathrm{MT}} \right\|_1 . \tag{1}$$

3.2 Stage 2: Model Distillation

In Stage 2, the network parameters of the optimized Knowledge Aggregation module will be fixed for student-teacher distillation. Here our new MTKD approach does not require the student network to have a similar architecture to one of the teacher models. This allows us to employ teacher models with diverse network architectures in order to achieve improved knowledge distillation performance.

Loss function The output of the Knowledge Aggregation module and the ground-truth image are jointly employed to train the student model. Specifically, we first compare the output of the student model I_{stu} with the ground truth based on L1 loss:

$$\mathcal{L}_{stu} = \left\| I_{\mathrm{stu}} - I_{\mathrm{GT}} \right\|_1 . \tag{2}$$

The design of the distillation loss is inspired by [76], where a wavelet-based training methodology was employed for image-to-image translation. It is calculated between I_{stu} and $I_{\mathrm{HR}}^{\mathrm{MT}}$ after decomposing both using a discrete wavelet transform (DWT):

$$\mathcal{L}_{dis} = \frac{1}{3K+1} \sum_{i,k} \left\| \mathrm{DWT}_{i,k}(I_{\mathrm{stu}}) - \mathrm{DWT}_{i,k}(I_{\mathrm{HR}}^{\mathrm{MT}}) \right\|_1 , \tag{3}$$

Table 1: The configurations of the employed teacher and student ISR networks.

Model	Role	Channel	Block/Group	#Params(M)
SwinIR [39]	Teacher	180	6/-	11.8
SwinIR_lightweight [39]	Student	60	4/-	0.9
RCAN [77]	Teacher	64	20/10	15.6
RCAN_lightweight [15, 78]	Student	64	6/10	5.2
EDSR [40]	Teacher	256	32/-	43
EDSR_baseline [40]	Student	64	16/-	1.5

in which $i \in \{\mathrm{LL}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}$. Here $k = 1 \ldots K$, which stands for the DWT decomposition level. The wavelet-based loss function used here is to allow the student model to enhance its ability by learning from its teachers through introducing high-frequency information.

These two losses are then combined as the overall loss function $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{stu}(I_{\mathrm{stu}}, I_{\mathrm{GT}}) + \mathcal{L}_{dis}(I_{stu}, I_{\mathrm{HR}}^{\mathrm{MT}}), \tag{4}$$

where α is a tunable weight to determine the contributions of two losses. It is noted here that we did not employ a wavelet-based loss for $\mathcal{L}_{stu}$. This is because (i) L1 loss is the most commonly used loss function to minimize the difference between the model output and the ground truth; (ii) due to the small α value used, the contribution of $\mathcal{L}_{stu}$ is rather limited and (iii) we observed training instability and performance reduction when wavelet-based loss is employed for $\mathcal{L}_{stu}$. Hence we followed common practice in [21, 23, 76] to use L1 loss here, which can effectively reduce the training complexity without compromising performance.

4 Experiment Configuration

Teacher and student networks To evaluate the effectiveness of the proposed MTKD approach, we followed the evaluation practice in [78] by selecting two widely used CNN-based ISR models: EDSR [40], RCAN [77], and one popular transformer-based model: SwinIR [39] for comparison. Their open-source full original models have been used here as the teacher networks. As the aim of this paper is to develop and validate a new knowledge distillation method, rather than design a novel lightweight model, we employed the existing low complexity variants of these teacher networks as the compact student models. For EDSR and SwinIR, the student models are the same as in their original papers: EDSR_baseline [40] and SwinIR_lightweight [39]. For RCAN, due to the lack of the lite model in the corresponding literature, we follow the configurations in [15, 78] to obtain the compact RCAN model, denoted by RCAN_lightweight. The configuration details of these teacher and student models are summarized in Table 1. To fully evaluate the ISR performance, we trained and evaluated each model for SR tasks with multiple scale factors, including $\times 2$, $\times 3$ and $\times 4$.

Training configurations In alignment with [39, 40, 77], we utilized 800 images from DIV2K [59] in this experiment for model training. In the training process, we perform random cropping of LR patches with dimensions 64×64 from the LR images, and the corresponding HR patches are cropped from the ground-truth images based on the scale factor. To achieve data augmentation, random rotation and horizontal flipping are further applied to the training material.

Other training configurations include ADAM optimizer [33] with parameter settings $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$; the block number B is 4, layer number L is 2 and the window size W_s is set to 8; the feature channel number C is 24, and the maximum DWT decomposition level K is 1; the factor α used to balance the distillation loss function is set as 0.1. The training batch size is 16 with a total of 2.5×10^5 iterations; the initial learning rate is set to 10^{-4} and is decayed by a factor of 10 at every 10^5 update. This experiment is implemented on the BasicSR [63] platform using an NVIDIA V100 GPU.

Evaluation configurations Four commonly used test sets, Set5 [3], Set14 [74], BSD100 [49], and Urban100 [24], were employed here for benchmarking the model performance. The ISR performance is assessed using two widely used quality metrics including peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [65].

To benchmark the performance of the proposed MTKD method, we employed five existing knowledge distillation methods for comparison including basic KD [23], AT [73], FAKD [21], DUKD [78], and CrossKD [15]. We also include the results of the corresponding pre-trained teacher and student (training from scratch without any KD) models for benchmarking.

5 Results and Discussion

5.1 Quantitative Evaluation

The quantitative results for three compact ISR networks, EDSR_baseline [40], SwinIR_lightweight [39] and RCAN_lightweight [15, 78] that have been trained with different knowledge distillation methods, including basic KD [23], AT [73], FAKD [21], DUKD [78], CrossKD [15] and the proposed MTKD are presented in Table 2-4, in which the best and second-best performers are highlighted in red and blue, respectively. Here we also provide results for the original EDSR, SwinIR and RCAN models (denoted by "full" in Table 2-4), their pre-trained student variants (without knowledge distillation, denoted by "compact"), and our enhanced teacher (through knowledge aggregation, denoted by "MT") for reference. It should be noted that the results for DUKD, CrossKD, and three original pre-trained teacher/student models are from their corresponding publications, while the results for KD, AT and FAKD were generated by ourselves based on their publicly available source code. We did not report the results of CrossKD for the EDSR and SwinIR student model, and those of DUKD for EDSR student,

Table 2: Quantitative results (PSNR/SSIM) for the RCAN_lightweight model.

		Dataset	Set5	Set14	BSD100	Urban100
Model	Scale	Method	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
RCAN	×2	MT	38.41/0.9626	34.49/0.9252	32.53/0.9045	33.91/0.9429
		Full	38.27/0.9614	34.12/0.9216	32.41/0.9027	33.34/0.9384
		Compact	38.07/0.9608	33.62/0.9183	32.20/0.9000	32.32/0.9302
		KD	38.18/0.9611	33.83/0.9197	32.29/0.9010	32.67/0.9329
		AT	38.13/0.9610	33.70/0.9187	32.25/0.9005	32.48/0.9313
		FAKD	38.16/0.9611	33.82/0.9190	32.27/0.9010	32.53/0.9320
		DUKD	38.23/0.9614	33.90/0.9201	32.33/0.9016	32.87/0.9349
		CrossKD	38.18/0.9612	33.82/0.9195	32.29/0.9012	32.69/0.9331
		MTKD (ours)	38.26/0.9619	34.09/0.9219	32.40/0.9031	33.06/0.9364
	×3	MT	34.96/0.9315	30.93/0.8531	29.48/0.8154	29.85/0.8822
		Full	34.74/0.9299	30.65/0.8482	29.32/0.8111	29.09/0.8702
		Compact	34.56/0.9284	30.41/0.8438	29.16/0.8076	28.48/0.8600
		KD	34.61/0.9291	30.47/0.8447	29.21/0.8080	28.62/0.8612
		AT	34.55/0.9287	30.43/0.8438	29.17/0.8070	28.43/0.8577
		FAKD	34.65/0.9291	30.45/0.8442	29.21/0.8087	28.52/0.8602
		DUKD	34.74/0.9296	30.54/0.8458	29.25/0.8088	28.79/0.8646
		CrossKD	34.66/0.9291	30.50/0.8448	29.22/0.8082	28.64/0.8617
		MTKD (ours)	34.78/0.9306	30.59/0.8483	29.34/0.8106	29.18/0.8704
	×4	MT	32.83/0.9027	29.06/0.7934	27.93/0.7496	27.44/0.8232
		Full	32.63/0.9002	28.87/0.7889	27.77/0.7436	26.82/0.8087
		Compact	32.32/0.8964	28.69/0.7840	27.63/0.7381	26.34/0.7933
		KD	32.45/0.8980	28.76/0.7860	27.67/0.7400	26.49/0.7980
		AT	32.31/0.8967	28.69/0.7839	27.64/0.7385	26.29/0.7927
		FAKD	32.46/0.8983	28.75/0.7859	27.68/0.7402	26.42/0.7973
		DUKD	32.56/0.8990	28.83/0.7870	27.72/0.7410	26.62/0.8020
		CrossKD	32.45/0.8984	28.81/0.7866	27.69/0.7406	26.53/0.7992
		MTKD (ours)	32.62/0.9009	28.84/0.7901	27.88/0.7447	27.08/0.8108

due to the unavailability of the source code and these results associated with their original papers.

It can be observed from Table 2-4 that our MTKD approach consistently offers the best performance for all three student models, on four different test datasets, and for different scale factors. MTKD optimized RCAN_lightweight model even outperforms its original full version (e.g., with a 0.26dB PSNR gain for scale ×4 on the Urban100 database), with only one-third of the model parameters. We also note that the performance improvement over other KD methods is more significant when the performance difference between the original teacher and (pre-trained w/o KD) student models is larger - this aligns with the observation reported elsewhere [51]. The second best performer is DUKD for RCAN

Table 3: Quantitative results (PSNR/SSIM) for the EDSR_baseline model.

		Dataset	Set5	Set14	BSD100	Urban100
Model	Scale	Method	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
EDSR	×2	MT	38.41/0.9626	34.49/0.9252	32.53/0.9045	33.91/0.9429
		Full	38.11/0.9601	33.92/0.9195	32.32/0.9013	32.93/0.9351
		Compact	37.96/0.9608	33.55/0.9176	32.17/0.9003	31.99/0.9274
		KD	37.97/0.9610	33.60/0.9180	32.19/0.9002	32.09/0.9283
		AT	37.99/0.9607	33.58/0.9173	32.21/0.8996	32.08/0.9275
		FAKD	38.01/0.9604	33.59/0.9181	32.23/0.9008	32.11/0.9292
		MTKD (ours)	38.08/0.9612	33.82/0.9196	32.29/0.9017	32.42/0.9308
	×3	MT	34.96/0.9315	30.93/0.8531	29.48/0.8154	29.85/0.8822
		Full	34.65/0.9282	30.52/0.8462	29.25/0.8093	28.80/0.8653
		Compact	34.36/0.9273	30.28/0.8421	29.09/0.8066	28.14/0.8528
		KD	34.39/0.9277	30.31/0.8427	29.10/0.8071	28.19/0.8533
		AT	34.40/0.9268	30.30/0.8431	29.16/0.8056	28.12/0.8519
		FAKD	34.47/0.9273	30.37/0.8425	29.12/0.8062	28.21/0.8533
		MTKD (ours)	34.54/0.9286	30.48/0.8450	29.20/0.8086	28.48/0.8578
	×4	MT	32.83/0.9027	29.06/0.7934	27.93/0.7496	27.44/0.8232
		Full	32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033
		Compact	32.09/0.8944	28.56/0.7814	27.57/0.7372	26.03/0.7849
		KD	32.12/0.8952	28.56/0.7823/	27.56/0.7382	26.02/0.7861
		AT	32.08/0.8935	28.49/0.7798	27.51/0.7365	26.00/0.7866
		FAKD	32.21/0.8957	28.55/0.7826	27.58/0.7377	26.11/0.7892
		MTKD (ours)	32.29/0.8967	28.73/0.7849	27.69/0.7407	26.32/0.7918

and FAKD for SwinIR and EDSR, typically with lower PSNR results (up to 0.46 dB) compared to MTKD. We also observe that most of the benchmarked knowledge distillation methods do offer improved performance compared to the corresponding pre-trained student model (w/o KD) - thus verifying the effectiveness of the knowledge distillation technique. As we mentioned above, the output of the Knowledge Aggregation module, $I_{\text{HR}}^{\text{MT}}$, has also been compared with its ground-truth counterpart, with results (denoted by "MT") are shown in Table 2-4. It is noted that for all three scale factors, and four test datasets, the MT results are always better than those for full EDSR, SwinIR and RCAN models, which also showcase the effectiveness of the Knowledge Aggregation module.

5.2 Qualitative Evaluation

Fig. 2 and 3 shows a visual comparison between results generated by various SwinIR models (for a scale factor 4) that are trained using different knowledge distillation methods. The source images presented are from the Urban100 dataset, which is the most challenging one among all four test sets. It can be

Table 4: Quantitative results (PSNR/SSIM) for the SwinIR_lightweight model.

		Dataset	Set5	Set14	BSD100	Urban100
Model	Scale	Method	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
SwinIR	×2	MT	38.41/0.9626	34.49/0.9252	32.53/0.9045	33.91/0.9429
		Full	38.35/0.9620	34.14/0.9227	32.44/0.9030	33.40/0.9393
		Compact	38.14/0.9611	33.86/0.9206	32.31/0.9012	32.76/0.9340
		KD	38.15/0.9619	33.90/0.9211	32.33/0.9023	32.79/0.9342
		FAKD	38.16/0.9615	33.87/0.9216	32.34/0.9024	32.81/0.9345
		AT	38.14/0.9615	33.85/0.9209	32.32/0.9021	32.74/0.9341
		DUKD	38.13/0.9610	33.78/0.9194	32.26/0.9007	32.63/0.9327
		MTKD (ours)	38.21/0.9619	34.03/0.9218	32.39/0.9030	32.92/0.9351
	×3	MT	34.96/0.9315	30.93/0.8531	29.48/0.8154	29.85/0.8822
		Full	34.89/0.9312	30.77/0.8503	29.37/0.8124	29.29/0.8744
		Compact	34.62/0.9289	30.54/0.8463	29.20/0.8082	28.66/0.8624
		KD	34.61/0.9292	30.55/0.8469	29.23/0.8100	28.67/0.8627
		FAKD	34.65/0.9296	30.56/0.8463	29.25/0.8113	28.65/0.8629
		AT	34.59/0.9289	30.53/0.8464	29.22/0.8094	28.63/0.8618
		DUKD	34.55/0.9285	30.53/0.8456	29.20/0.8080	28.53/0.8604
		MTKD (ours)	34.70/0.9300	30.66/0.8480	29.31/0.8116	29.03/0.8679
	×4	MT	32.83/0.9027	29.06/0.7934	27.93/0.7496	27.44/0.8232
		Full	32.72/0.9021	28.94/0.7914	27.83/0.7459	27.07/0.8164
		Compact	32.44/0.8976	28.77/0.7858	27.69/0.7406	26.47/0.7980
		KD	32.43/0.8984	28.78/0.7862	27.71/0.7426	26.50/0.7981
		FAKD	32.43/0.8979	28.79/0.7857	27.69/0.7427	26.51/0.7982
		AT	32.42/0.8982	28.77/0.7860	27.70/0.7421	26.47/0.7981
		DUKD	32.41/0.8973	28.79/0.7860	27.69/0.7405	26.43/0.7972
		MTKD (ours)	32.52/0.8993	28.87/0.7885	27.79/0.7450	26.85/0.8071

observed that for repetitive texture reconstruction (subfigure (a) and (b)) in Fig. 2, MTKD produces sharper edges and structures which are closer to the ground-truth high-resolution images compared to other knowledge distillation methods. This may be due to the use of the DWT-based loss function, which is able to perceive high frequency energy changes at various directions. MTKD can also generate more spatial details as shown in subfigures (c) and (d) in Fig. 3, such as complex textures. The perceptual quality enhancement achieved by MTKD has also been verified by the corresponding PSNR values in each example.

5.3 Ablation Study

To demonstrate the effectiveness of each primary contribution in this work, we conducted ablation studies to more fully characterise the impact of our proposed MTKD framework.

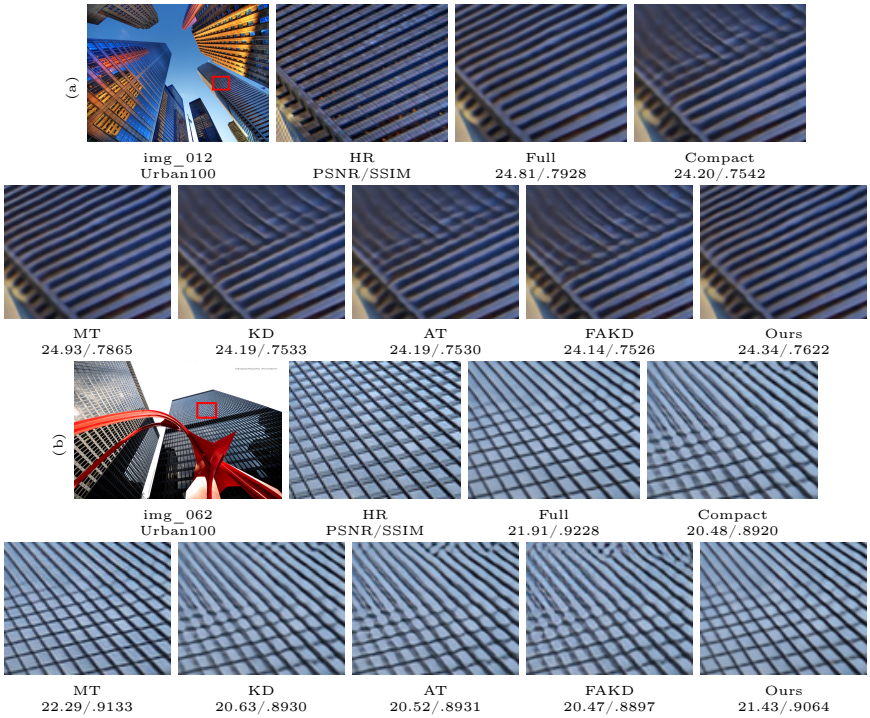

Fig. 2: The ×4 super-resolution results of **SwinIR** models on (a) img012, (b) img062 from Urban100. PSNRs and SSIMs are displayed below each image.

Study 1: Multiple teachers To confirm the contribution of multiple teachers in the MTKD framework, we created two different variants, each of which employs 1-2 teachers. Here only the RCAN_lightweight model (for the ×4 ISR task) is employed in this experiment, and the created variants include (v1) with both SwinIR and EDSR as teachers; (v2) only with SwinIR as teacher. Here we kept SwinIR as a teacher in both variants due to its superior performance over the other two teachers. As shown in Table 5, the performance is improved in line with the number of teachers, with the full MTKD method (with three teachers) delivering the best ISR results.

To further showcase the contribution of each employed teacher model in the proposed framework, we utilize the Local Attribution Maps tool [19] to identify the level of contribution from the pixels in each teacher's output, I^{n}_{HR}, to those in the output of the Knowledge Aggregation module, $I^{\mathrm{MT}}_{\mathrm{HR}}$. As illustrated in Fig. 4, we can observe that all three teacher models have contributed information to the final output of the Knowledge Aggregation module.

Study 2: Knowledge aggregation network structure As the design of the Knowledge Aggregation network (with the new DCTSwin blocks) is one of the

Table 5: Ablation study results with RCAN_lightweight (×4) as the student model.

Variants	v1	v2	v3	v4
PSNR/SSIM	26.79/0.8018	26.70/0.8003	26.93/0.8062	26.89/0.8043
Variants	v5	v6	v7	**Ours**
PSNR/SSIM	26.78/0.8023	26.98/0.8078	26.28/0.7958	27.08/0.8108

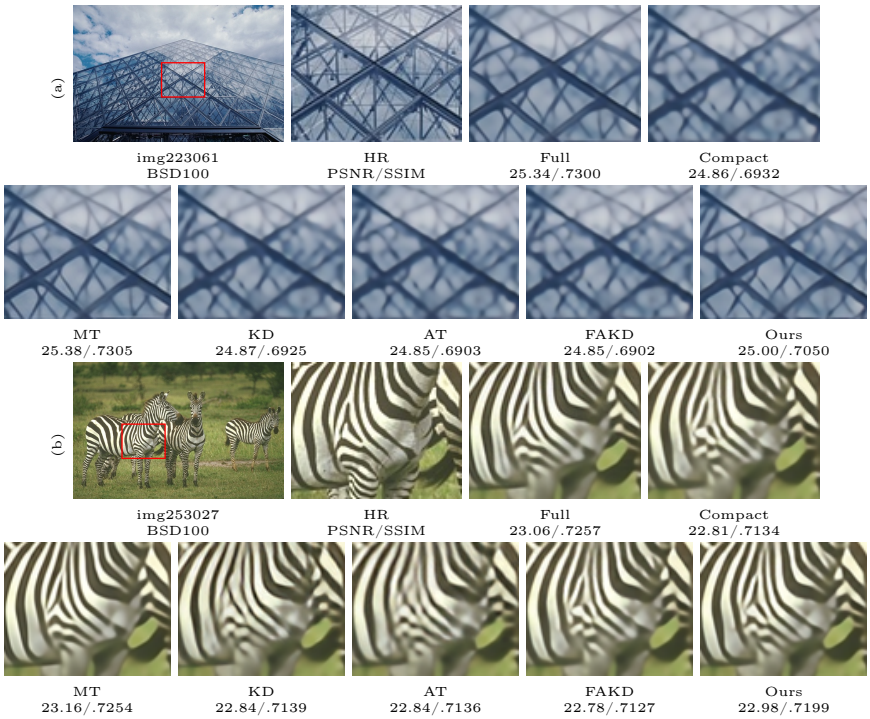

Fig. 3: The ×4 super-resolution results of **SwinIR** models on (a) img223061 and (b) img253027 from BSD100. PSNRs and SSIMs are displayed below each image.

primary contributions in this work, we tested the effectiveness of the proposed DCTSwin blocks by replacing them with Mixer Layer blocks [60] (v3) and removing the DCT and IDCT modules (v4). In (v3), to enable a fair comparison, we keep the number of parameters similar to that of the original knowledge aggregation network. The experiment is also based on the Urban100 database, RCAN_lightweight model and the scale factor 4. Based on the results summarized in Table 5, both (v3) and (v4) have been outperformed by the original network design, which shows the importance of using DCT/IDCT modules and the DCTSwin blocks.

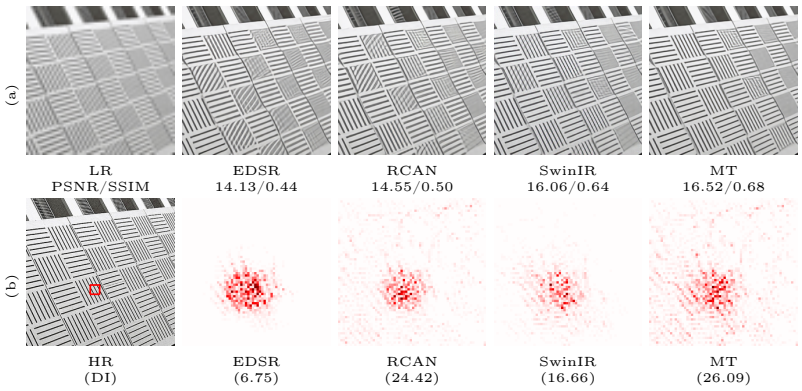

Fig. 4: The illustration of the contribution from each teacher model using the Local Attribution Maps tool [19]. (a) The input low-resolution image and the high resolution reconstructions generated by the original EDSR, RCAN, and SwinIR at a $\times 4$ scale. (b) the high-resolution image and LAM [19] maps for all three teacher models. Here diffusion Index (DI) quantifies the overall contribution from each teacher.

Study 3: Distillation loss function To validate the effectiveness of the wavelet-based distillation loss, we have replaced it with L1 loss (v5), DCT-based loss [61] (v6) and an alternative DWT loss [76] (which only focuses on high-frequency information) (v7) but kept the network architectures and training configurations the same. This study is based on the Urban100 dataset and the RCAN_lightweight model for the $\times 4$ image super-resolution task. The results shown in Table 5 indicate that the employed wavelet-based distillation loss does improve the model performance compared to other tested loss functions.

6 Conclusion

This paper presents a novel Multi-Teacher Knowledge Distillation (MTKD) framework for image super-resolution. The proposed approach integrates a new DCTSwin-based network to aggregate the knowledge from multiple teacher models and generates an enhanced representation of the high-resolution image. This is employed to optimize the student model through distillation using a loss function based on discrete wavelet transform. We conduct comprehensive experiments for the image super-resolution task using various teacher and student networks and diverse test databases. Throughout these experiments, our method consistently outperforms other existing knowledge distillation methods, showcasing its effectiveness and robustness for the image super-resolution task. The primary contributions of this work have also been verified in the additional ablation study. Future work should focus on the application of this approach to other low-level computer vision tasks and different network architectures, e.g., diffusion models.

Acknowledgements

The authors appreciate the funding from Netflix Inc., the University of Bristol, and the UKRI MyWorld Strength in Places Programme (SIPF00006/1).

References

1. Afonso, M., Zhang, F., Bull, D.R.: Video compression based on spatio-temporal resolution adaptation. IEEE Transactions on Circuits and Systems for Video Technology **29**(1), 275–280 (2018)
2. Aleissaee, A.A., Kumar, A., Anwer, R.M., Khan, S., Cholakkal, H., Xia, G.S., Khan, F.S.: Transformers in remote sensing: A survey. Remote Sensing **15**(7), 1860 (2023)
3. Bevilacqua, M., Roumy, A., Guillemot, C., Morel, M.L.A.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: British Machine Vision Conference (BMVC) (2012)
4. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006)
5. Bull, D., Zhang, F.: Intelligent image and video compression: communicating pictures. Academic Press (2021)
6. Cai, J., Meng, Z., Ding, J., Ho, C.M.: Real-time super-resolution for real-world images on mobile devices. In: 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 127–132. IEEE (2022)
7. Chawla, A., Yin, H., Molchanov, P., Alvarez, J.: Data-free knowledge distillation for object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3289–3298 (2021)
8. Chebotar, Y., Waters, A.: Distilling knowledge from ensembles of neural networks for speech recognition. In: Interspeech. pp. 3439–3443 (2016)
9. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. Advances in neural information processing systems **30** (2017)
10. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5008–5017 (2021)
11. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22367–22377 (June 2023)
12. Chira, D., Haralampiev, I., Winther, O., Dittadi, A., Liévin, V.: Image super-resolution with deep variational autoencoders. In: European Conference on Computer Vision. pp. 395–411. Springer (2022)
13. Conde, M.V., Choi, U.J., Burchi, M., Timofte, R.: Swin2SR: Swinv2 transformer for compressed image super-resolution and restoration. In: European Conference on Computer Vision. pp. 669–687. Springer (2022)
14. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
15. Fang, H., Hu, X., Hu, H.: Cross knowledge distillation for image super-resolution. In: Proceedings of the 2022 6th International Conference on Video and Image Processing. pp. 162–168 (2022)

16. Feng, C., Danier, D., Wang, H., Zhang, F., Vallade, B., Mackin, A., Bull, D.: Rankdvqa-mini: Knowledge distillation-driven deep video quality assessment. In: 2024 Picture Coding Symposium (PCS). pp. 1–5. IEEE (2024)
17. Gao, Q., Zhao, Y., Li, G., Tong, T.: Image super-resolution using knowledge distillation. In: Asian Conference on Computer Vision. pp. 527–541. Springer (2018)
18. Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., Zhang, B.: Implicit diffusion models for continuous super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10021–10030 (2023)
19. Gu, J., Dong, C.: Interpreting super-resolution networks with local attribution maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9199–9208 (2021)
20. Gu, Y., Deng, C., Wei, K.: Class-incremental instance segmentation via multi-teacher networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1478–1486 (2021)
21. He, Z., Dai, T., Lu, J., Jiang, Y., Xia, S.T.: FAKD: Feature-affinity based knowledge distillation for efficient image super-resolution. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 518–522. IEEE (2020)
22. Heil, C.E., Walnut, D.F.: Continuous and discrete wavelet transforms. SIAM review **31**(4), 628–666 (1989)
23. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
24. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5197–5206 (2015)
25. Jacob, G.M., Agarwal, V., Stenger, B.: Online knowledge distillation for multi-task learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2359–2368 (2023)
26. Jeevan, P., Srinidhi, A., Prathiba, P., Sethi, A.: WaveMixSR: Resource-efficient neural network for image super-resolution. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5884–5892 (2024)
27. Jeevan, P., Viswanathan, K., Sethi, A.: WaveMix: A resource-efficient neural network for image analysis. arXiv preprint arXiv:2205.14375 (2022)
28. Jiang, Y., Nawała, J., Zhang, F., Bull, D.: Compressing deep image super-resolution models. In: 2024 Picture Coding Symposium (PCS). pp. 1–5. IEEE (2024)
29. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351 (2019)
30. Jin, Y., Wang, J., Lin, D.: Multi-level logit distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24276–24285 (2023)
31. Kang, J., Xu, W., Ritter, A.: Distill or annotate? cost-efficient fine-tuning of compact models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2023)
32. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
33. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

34. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
35. Lee, W., Lee, J., Kim, D., Ham, B.: Learning with privileged information for efficient image super-resolution. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. pp. 465–482. Springer (2020)
36. Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S.: Fnet: Mixing tokens with fourier transforms. arXiv preprint arXiv:2105.03824 (2021)
37. Lepcha, D.C., Goyal, B., Dogra, A., Goyal, V.: Image super-resolution: A comprehensive review, recent trends, challenges and applications. Information Fusion **91**, 230–260 (2023)
38. Li, W.H., Bilen, H.: Knowledge distillation for multi-task learning. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 163–176. Springer (2020)
39. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
40. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
41. Lin, S., Xie, H., Wang, B., Yu, K., Chang, X., Liang, X., Wang, G.: Knowledge distillation via the target-aware transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10915–10924 (2022)
42. Lin, Z., Garg, P., Banerjee, A., Magid, S.A., Sun, D., Zhang, Y., Van Gool, L., Wei, D., Pfister, H.: Revisiting rcan: Improved training for image super-resolution. arXiv preprint arXiv:2201.11279 (2022)
43. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023)
44. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
45. Liu, Z.S., Siu, W.C., Chan, Y.L.: Photo-realistic image super-resolution via variational autoencoders. IEEE Transactions on Circuits and Systems for video Technology **31**(4), 1351–1365 (2020)
46. López-Cifuentes, A., Escudero-Viñolo, M., Bescós, J., SanMiguel, J.C.: Attention-based knowledge distillation in multi-attention tasks: The impact of a DCT-driven loss. arXiv preprint arXiv:2205.01997 (2022)
47. Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., Zeng, T.: Transformer for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 457–466 (2022)
48. Luo, S., Wang, X., Fang, G., Hu, Y., Tao, D., Song, M.: Knowledge amalgamation from heterogeneous networks by common feature learning. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI) (2019)
49. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 2, pp. 416–423. IEEE (2001)

50. Meng, Z., Yao, X., Sun, L.: Multi-task distillation: Towards mitigating the negative transfer in multi-task learning. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 389–393. IEEE (2021)
51. Miles, R., Mikolajczyk, K.: Understanding the role of the projector in knowledge distillation. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI-24) (December 2023)
52. Morris, C., Danier, D., Zhang, F., Anantrasirichai, N., Bull, D.R.: ST-MFNet Mini: Knowledge distillation-driven frame interpolation. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 1045–1049 (2023)
53. Park, J., Son, S., Lee, K.M.: Content-aware local gan for photo-realistic super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10585–10594 (2023)
54. Peng, T., Gao, G., Sun, H., Zhang, F., Bull, D.: Accelerating learnt video codecs with gradient decay and layer-wise distillation. In: 2024 Picture Coding Symposium (PCS). pp. 1–5. IEEE (2024)
55. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
56. Shi, W., Ren, G., Chen, Y., Yan, S.: ProxylessKD: Direct knowledge distillation with inherited classifier for face recognition. arXiv preprint arXiv:2011.00265 (2020)
57. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
58. Singh, G., Mittal, A.: Various image enhancement techniques-a critical review. International Journal of Innovation and Scientific Research **10**(2), 267–274 (2014)
59. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: NTIRE 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 114–125 (2017)
60. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in neural information processing systems **34**, 24261–24272 (2021)
61. Tomosada, H., Kudo, T., Fujisawa, T., Ikehara, M.: Gan-based image deblurring using DCT loss with customized datasets. IEEE Access **9**, 135224–135233 (2021)
62. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
63. Wang, X., Xie, L., Yu, K., Chan, K.C., Loy, C.C., Dong, C.: BasicSR: Open source image and video restoration toolbox. `https://github.com/XPixelGroup/BasicSR` (2022)
64. Wang, Z., Chen, J., Hoi, S.C.: Deep learning for image super-resolution: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3365–3387 (2020)
65. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)

66. Wu, M.C., Chiu, C.T., Wu, K.H.: Multi-teacher knowledge distillation for compressed video action recognition on deep neural networks. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2202–2206. IEEE (2019)
67. Xu, R., Kang, X., Li, C., Chen, H., Ming, A.: DCT-FANet: DCT based frequency attention network for single image super-resolution. Displays **74**, 102220 (2022)
68. Yadav, O., Ghosal, K., Lutz, S., Smolic, A.: Frequency-domain loss function for deep exposure correction of dark images. Signal, Image and Video Processing **15**(8), 1829–1836 (2021)
69. Yang, J., Huang, T.: Image super-resolution: Historical overview and future challenges. In: Super-resolution imaging, pp. 1–34. CRC Press (2017)
70. Yao, G., Li, Z., Bhanu, B., Kang, Z., Zhong, Z., Zhang, Q.: MTKDSR: Multi-teacher knowledge distillation for super resolution image reconstruction. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 352–358. IEEE (2022)
71. Ye, J., Ji, Y., Wang, X., Ou, K., Tao, D., Song, M.: Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2829–2838 (2019)
72. Yuan, F., Shou, L., Pei, J., Lin, W., Gong, M., Fu, Y., Jiang, D.: Reinforced multi-teacher selection for knowledge distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14284–14291 (2021)
73. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017), https://arxiv.org/abs/1612.03928
74. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7. pp. 711–730. Springer (2012)
75. Zhang, D., Huang, F., Liu, S., Wang, X., Jin, Z.: SwinFIR: Revisiting the swinir with fast Fourier convolution and improved training for image super-resolution. arXiv preprint arXiv:2208.11247 (2022)
76. Zhang, L., Chen, X., Tu, X., Wan, P., Xu, N., Ma, K.: Wavelet knowledge distillation: Towards efficient image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12464–12474 (2022)
77. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)
78. Zhang, Y., Li, W., Li, S., Hu, J., Chen, H., Wang, H., Tu, Z., Wang, W., Jing, B., Wang, Y.: Data upcycling knowledge distillation for image super-resolution. arXiv preprint arXiv:2309.14162 (2023)
79. Zhu, X., Gong, S., et al.: Knowledge distillation by on-the-fly native ensemble. Advances in neural information processing systems **31** (2018)
80. Zhu, Y., Zhou, Q., Liu, N., Xu, Z., Ou, Z., Mou, X., Tang, J.: ScaleKD: Distilling scale-aware knowledge in small object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19723–19733 (2023)