IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 November 2023, accepted 18 November 2023, date of publication 23 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3335934

RESEARCH ARTICLE

Feature-Domain Adaptive Contrastive Distillation for Efficient Single Image Super-Resolution

HYEON-CHEOL MOON[1,2], JAE-GON KIM[2], (Member, IEEE), JINWOO JEONG[1], AND SUNGJEI KIM[1]

[1]Korea Electronics Technology Institutes (KETI), Seongnam 13509, Republic of Korea
[2]School of Electronics and Information Engineering, Korea Aerospace University, Goyang 10540, Republic of Korea

Corresponding authors: Sungjei Kim (sungjei.kim@keti.re.kr) and Jinwoo Jeong (jw.jeong@keti.re.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korean Government [Ministry of Science and Information and Communication Technology (MSIT)] (Development of Intelligent Media Aspect Ratio Conversion Technology That Maintains Properties) under Grant 2021-0-00802.

ABSTRACT Convolutional neural network-based single image super-resolution (SISR) involves numerous parameters and high computational expenses to ensure improved performance, limiting its applicability in resource-constrained devices such as mobile phones. Knowledge distillation (KD), which transfers useful knowledge from a teacher network to a student network, has been investigated as a method to make networks more efficient in terms of performance. To this end, feature distillation (FD) has been utilized in KD to minimize the Euclidean distance-based loss of feature maps between teacher and student networks. However, this technique does not adequately consider the effective and meaningful delivery of knowledge from the teacher to the student network to improve the latter's performance under given network capacity constraints. In this study, we propose a feature-domain adaptive contrastive distillation (FACD) method to train lightweight student SISR networks efficiently. We highlight the limitations of existing FD methods in terms of Euclidean distance-based loss, and propose a feature-domain contrastive loss, which causes student networks to learn richer information from the teacher's representation in the feature domain. We also implement adaptive distillation that performs distillation selectively depending on the conditions of the training patches. Experimental results demonstrated that the proposed FACD scheme improves student enhanced deep residual networks and residual channel attention networks not only in terms of the peak signal-to-noise ratio (PSNR) on all benchmark datasets and scales but also in terms of subjective image quality, compared to the conventional FD approaches. In particular, FACD achieved an average PSNR improvement of 0.07 dB over conventional FD in both networks. Code will be release at https://github.com/hcmoon0613/FACD.

INDEX TERMS Contrastive learning, efficient super-resolution, feature distillation, knowledge distillation, single image super-resolution.

I. INTRODUCTION

Single image super-resolution (SISR) is a method of generating a high-resolution image from a given low-resolution image [1]. It is an important method that can be applied to a variety of computer vision tasks, such as medical imaging [2], satellite [3], [4], [5], remote sensing [6], [7], [8], face hallucination [9], [10], [11], and object recognition [12], [13]. In prior works, interpolation and example-based methods have been applied for SISR [14], [15], [16]. However, both approaches exhibit performance limitations. Recently, convolutional neural network (CNN)-based SISR networks, such as SRCNN [17], have been reported to outperform traditional SISR works. Since then, numerous CNN-based SISR networks have been proposed [1], [18], [19], and the network parameters and computational complexity have been increased to obtain better performance.

The practical applicability of complex SISR models is limited in resource-constrained devices, such as mobile or IoT devices; thus, efficient and lightweight SISR models are

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

FIGURE 1. Examples of limitations on feature distillation (FD) with Euclidean loss [20], [21]. The teacher and student networks are student enhanced deep residual networks (EDSR) [22] with scale x2 image. Note that feature affinity-based knowledge distillation (FAKD*) and local feature distillation (LFD*) indicate our reproduced results with the same experimental settings.

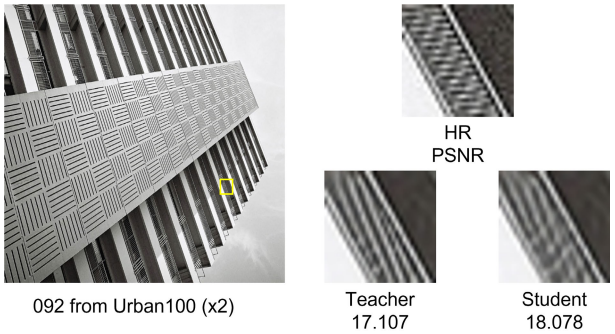

FIGURE 2. Examples of worse teacher case on enhanced deep residual network (EDSR) network with KD scheme (scale 2). Note that teacher and student networks are trained separately from scratch.

required. To satisfy this demand, lightweight SISR models with a better trade-off between efficiency and performance quality have been proposed [20], [21], [23], [24], [25], [26], [27], [28]. Among the aforementioned methods, knowledge distillation (KD) [29]-based approaches exhibit the following distinctive advantages: 1) KD promotes the inheritance of the knowledge of large teacher networks and improves performance without modifying the existing network structure at industrial sites, and 2) KD can be combined with pruning and network design methods by including additional loss terms to achieve greater performance improvement [30], [31].

KD is primarily used for classification and detection tasks [12], [32], [33]. The student network is trained to minimize the distance between the labels of the student network and the soft labels of the teacher network in the classification task. However, this approach to SISR exhibits limited performance improvement [21], [34]. To solve this problem, feature distillation (FD) is used to guide the training of the student network. Feature affinity-based KD (FAKD) [20] transfers the intermediate feature knowledge of a larger teacher model to a lightweight student network. In FAKD, FD with image domain distillation improves the distillation performance. Subsequently, local feature distillation/local-selective feature distillation (LFD/LSFD) [21] has been proposed based on the feature attention method, which selectively focuses on specific positions to extract refined feature information, improving the simple distance-based feature distillation of FAKD. Both methods use Euclidean distance as a metric to transfer feature knowledge from teacher to student networks. As depicted in Fig. 1, neither method eliminates the disadvantages of Euclidean distance-based loss, such as pattern loss and image blurring.

To overcome the limitations of Euclidean distance loss, contrastive distillation for SISR has been studied in the context of the KD scheme [25], [35]. Contrastive self-distillation (CSD) [25] explicitly transfers knowledge from teacher to student networks using contrastive loss in the latent space of the image domain and improves distillation performance and texture restoration. However, in SISR, this approach may degrade distillation efficiency because it cannot completely leverage the rich information of intermediate feature maps. Furthermore, CSD suffers from unstable distillation performance owing to its use of external images as negative samples for contrastive loss.

Finally, as depicted in Fig. 2, the inference output of a teacher network does not guarantee better performance for all patches. Inappropriate inference results interfere with the training of the student network and must be removed from the training process.

To address these problems, we propose feature-domain adaptive contrastive distillation (FACD), which selectively transfers the teacher's feature-domain knowledge using contrastive loss. The proposed feature-domain contrastive distillation (FCD) resolves the restoration of edges and patterns, and improves distillation performance compared to image-domain contrastive distillation (ICD) and CSD [25] by transferring well-refined feature knowledge to the student network effectively.

In addition, FAKD and LSFD, which use three intermediate feature maps for FD, do not account for attention at the feature map level. As CNN-based super-resolution (SR) networks have a cascading structure, improper distillation at the upper part has been found to sequentially affect the output at the lower part of the network. To this end, we assign greater importance (attention) to the top of network, as discussed in Section V.

Finally, the percentage of inappropriate inference outputs of the teacher network amounts to up to 5% for enhanced deep residual networks (EDSR) [22] and up to 11% for residual channel attention networks (RCAN) [36] across all training patches. Because these patches can transfer incorrect knowledge to the student network, the application of FCD is adaptively adjusted based on patch conditions during training. Combined with feature-domain contrastive loss, feature map level attention, and adaptive distillation, FACD achieves state-of-the-art performance over FD and excellent qualitative results. Our main contributions can be summarized as follows:

1) We propose an algorithm called FCD that improves the efficiency of traditional FD and mitigates the loss of detailed texture information that occurs in

the FD method based on an intermediate feature domain contrastive learning method, which refines and transfers useful representational knowledge to students.
2) We observe that inappropriate teacher network knowledge interferes with student network learning. To ensure efficient distillation, we propose an algorithm called FACD that selectively applies FCD by comparing output patches derived from teacher and student networks with the ground truth.
3) We demonstrate that the FACD achieves state-of-the-art performance compared to distance-based FD and yields excellent qualitative results. In particular, FACD outperforms FD in terms of edge and pattern reconstruction results. Furthermore, we conduct extensive experimental analysis with ablation studies.

The remainder of this paper is organized as follows. In Section II, we discuss related works on super-resolution. The proposed method is described in Section III. The effectiveness of the FACD is evaluated in Section IV and V. Finally, the conclusions of this study are summarized in Section VI.

II. RELATED WORKS

A. EFFICIENT SUPER-RESOLUTION NETWORK

At first, CNN-based SR models stack deeper layers to improve performance; however, this induces gradient vanishing. Subsequently, very-deep SR [18] and deeply recursive convolutional [19] networks have been proposed, which use deep stacking of residual blocks [37] to solve this problem. In addition, batch normalization (BN) has been applied to the SISR model in EDSR [22], thereby normalizing the features and eliminating model flexibility. EDSR uses the residual-scaling method to improve training instability induced by the removal of BN. Furthermore, residual dense network (RDN) fully exploit the features from all layers for utilizing hierarchical features. Especially, RDN used the residual dense block (RDB) which adaptively learn more preceding and current local features for stable training [38]. After that, residual channel attention networks (RCAN) [36] and second-order attention networks (SAN) [39] have been reported to achieve significant performance improvements by adopting the channel-attention mechanism. Since then, fast and memory efficient network (FMEN) has adopted the sequential attention branch, in which spatial pixel is assigned an important factor according to local and global contexts [40]. Recently, multi-level dispersion residual network (MDRN) achieved the first place in the NTIRE 2023 Efficient SR Challenge by adopting the attention distillation and multi-level dispersion spatial attention mechanism [41]. On the other hand, generative adversarial network (GAN) based SISR attempt to generate perceptual texture through learning with adversarial loss [42], [43]. Recent GAN-based SISR have achieved significant performance gains by utilizing the rich and diverse priors encapsulated in pre-trained GAN models for adversarial loss [44], [45].

However, the utilization of deep layers, stacked blocks, and attention mechanisms results in substantial memory and computational expenses during inference due to the considerable parameter count and the execution of spatial and non-local operations. Moreover, their applicability is limited on resource-constrained devices, such as mobile phones or IoT devices. To adapt SISR to such devices, the development of an efficient network structure and the optimization of training schemes is essential [46]. Because performance optimization achieved by only designing an efficient network structure is limited, advanced training schemes, comprising pruning, quantization, and KD, are very important on resource-constrained devices. Among these, KD is particularly promising because it achieves additional performance improvement without requiring the structure of the target model to be changed. This approach is described in detail in the next section.

B. FEATURE-DOMAIN DISTILLATION FOR SISR

In KD, knowledge is transferred from a teacher model to lightweight student networks [29]. Distillation based on the label domain (identical to the image domain in SISR) yields better classification performance. However, in regression problems such as SISR, the solution space is very large; therefore, single image-domain KD is not an effective method to transfer knowledge [47]. Therefore, FD was proposed to guide the training of the student network effectively based on Euclidean distance-based matching in the image and feature domains [20], [21], [34], [48].

First, FitNet [34] was proposed based on distillation in both the image and feature domains. For FD, a simple regressor composed of 1×1 convolution layers was used owing to differences in the channel size of the teacher and student networks. On the other hand, PISR [48] has been proposed, in which ground truth images are used as privileged information to teach an encoder in the teacher network the degradation and sub-sampling of high-resolution images. For more efficient distillation, FAKD is a feature affinity matrix-based KD framework that distils the structural knowledge from a larger teacher model. Furthermore, the teacher supervision (TS) loss between the output SR images of teacher and student networks is considered. In addition, LSFD is a feature attention method that adaptively focuses on specific pixels to extract feature information using the difference map between the inference output of teacher and student networks. By merging FD with the adaptive functional attention mechanism, LSFD exhibits enhanced performance compared to other FD algorithms such as FAKD. However, these approaches do not completely address the limitation of the Euclidean distance loss in terms of performance and subjective image quality. FD may degrade distillation efficiency because it cannot completely distill the rich information of intermediate feature maps. Therefore, in this paper, we propose a feature-domain contrastive loss, which causes student networks to transfer richer information from the teacher's representation in

the feature domain by maximizing mutual information. We will describe the details of contrastive loss in the next subsection.

C. CONTRASTIVE LEARNING

Contrastive loss is introduced in self-supervised learning [49], [50] and it is employed to train images to ensure that positive pairs remain close to each other, while negative pairs remain far away [25], [32]. By maximizing the Kullback-Leibler (KL) divergence of the positive and negative pairs, the mutual information in the positive pair is maximized, while both distributions are clearly distinguished. In other words, KD with contrastive loss optimizes performance by maximizing mutual information between teacher and student networks, while minimizing uncertainty between both networks simultaneously; thus, as training progresses, teacher and student networks become gradually similar. In this approach, contrastive representation distillation (CRD) achieves the better results than convectional KD in classification tasks by distillation based on contrastive loss [32]. Recently, complementary relation contrastive distillation (CRCD) distilled the relation structural knowledge between teachers and students to achieve better performance than CRD, a sample-based contrastive distillation [51].

Similar to the classification task, contrastive loss-based methods in SISR or restoration have led to performance improvements over conventional methods [52], [53], [54], [55]. Most of studies have performed contrastive learning by generating positive and negative pairs from the output images (or feature embedding) of the SR networks, and have shown effectiveness in terms of texture restoration. As a results, KD with contrastive loss [25], [35] in the image domain has been proposed for SISR, resulting in a marginal performance enhancement compared to other distillation approaches [20], [21]. In particular, CSD [25] uses contrastive loss in the latent features of the image domain. Inspired by conventional FD and CSD, we focus on improving distillation by using contrastive loss in the feature domain, where the solution space in the regression task is smaller than in the image domain. Therefore, we propose a novel method of feature-domain contrastive distillation and introduce an adaptive KD approach for efficient knowledge transfer from teacher networks.

III. PROPOSED METHOD

In this section, we describe an overall architecture and the loss function of the proposed distillation method. The pipeline of the proposed FACD framework is depicted in Fig. 3. The proposed FACD performs distillation in both image and feature domains. In the image domain, the output images of the teacher network and ground truth (GT) are used for KD of the student network. On the other hand, in the feature domain, FACD operates based on contrastive learning between the intermediate feature maps of the teacher and student networks [20], [21].

A. ADAPTIVE KNOWLEDGE DISTILLATION FOR SR

In this section, we describe adaptive KD for efficient knowledge transfer from teacher networks. As depicted in Fig. 2, the output of the teacher network is not always guaranteed to have better performance than student networks. On average, EDSR and RCAN yield 5% and 11% worse cases, respectively, on the training patches, which interfere with the efficiency of distillation. Therefore, we propose a simple but effective adaptive distillation method to optimize the distillation performance in both the image and feature domains. If the SR image of the student network is closer to the ground truth than the SR image of the teacher network, we ignore these patches during training. The indicator of adaptive KD is formulated as follows:

$$\alpha_i = \begin{cases} 0 & \text{if } \|SR_i^S - GT_i\|_1 < \|SR_i^T - GT_i\|_1, \\ 1 & \text{else.} \end{cases} \tag{1}$$

where α denotes the indicator of appropriate samples, and i denotes the index of the batch sample. If the distance from GT is farther from the teacher, the parameters of appropriate samples α_i are set to 0, indicating that the patch is not used for distillation.

B. CONTRASTIVE ADAPTIVE DISTILLATION

Conventional FD has demonstrated that the efficiency of distillation lies in the feature domain rather than the image domain [20]. Consequently, previous FD methods for SISR have focused primarily on improving distillation schemes in the feature domain. Nevertheless, the combination of feature distillation and image distillation has been shown to outperform a single refined feature distillation [20], [21]. Furthermore, KD with contrastive loss improves performance by enabling more explicit knowledge transfer from teacher networks [25]. Therefore, we apply contrastive learning in the feature domain to transfer richer information from intermediate features, and retain the Euclidean distance-based loss during image domain distillation to minimize interference induced by contrastive learning. We describe the loss function for each domain separately.

First, we propose the loss function in the image domain with the adaptive KD indicator (α_i) as follows:

$$L_{SR} = \frac{1}{2N}\sum_{i=1}^{N}(2-\alpha_i)\|SR_i^S - GT_i\|_1 + \alpha_i\|SR_i^S - SR_i^T\|_1 \tag{2}$$

where SR^S, SR^T, and GT denote the output images of the student network, teacher network, and GT, respectively. N indicates the batch size of training process. L_{SR} denotes the combined loss function of the Euclidean loss with GT (conventional SISR) and the distillation loss in the image domain.

Second, for the purpose of transferring knowledge from intermediate features of teacher networks, we propose feature domain adaptive contrastive distillation (FACD). To ensure

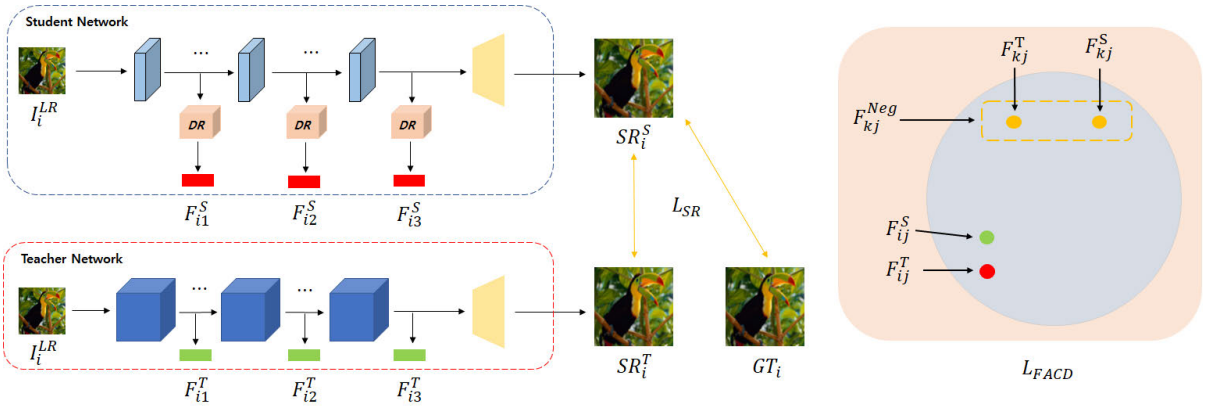

FIGURE 3. Overall architecture of feature-domain adaptive contrastive distillation (FACD).

fair comparison with conventional FD methods [20], [21], FACD configures three feature matching points, as depicted in Fig. 3. The detailed loss function of FACD is formulated as follows:

$$\hat{F}_{ij} = \frac{F_{ij}}{\|F_{ij}\|_2} \tag{3}$$

$$L_{FACD} = \sum_{i=1}^{N}\sum_{j=1}^{3} w_j \frac{\alpha_i \|DR(\hat{F}_{ij}^S) - \hat{F}_{ij}^T\|_1}{\sum_{k=1}^{K} \|DR(\hat{F}_{ij}^S) - \hat{F}_{kj}^{Neg}\|_1} \tag{4}$$

where $\hat{F}_{ij}$ denotes the normalized feature maps, w_j represents the attention weight of each feature matching point, and DR denotes the deep regressor comprising five 1×1 convolutional layers with PReLU activation [56]. In this study, the feature map level attention parameter w_i was set to [0.5, 0.3, 0.2]. In addition, N refers to the number of batch sizes, and K refers to the number of negative pairs. To transfer knowledge from the teacher network more effectively, the feature maps are first normalized. As given by Eq. 2, FD is not performed on inappropriate samples in positive pairs.

To use contrastive loss on the feature domain, the construction method of positive and negative pairs, as well as the similarity measures (e.g., Euclidean distance, dot-product, or cosine similarity) to be used in the contrastive loss function should be determined.

As depicted in Fig. 3, for the proposed FACD loss, we consider the features of the student network F_{ij}^S and the teacher network F_{ij}^T as a positive pair in the same index. To enhance the efficiency of contrastive distillation, as depicted in Fig. 3, all features of teacher and student networks except the ones corresponding to the same index are considered as negative pairs. To generate more negative samples, pairs with different indices on the student's features (F_{ij}^S, F_{kj}^S) are also considered as negative pairs. Moreover, contrastive loss with Euclidean loss is adopted as the similarity measure.

By minimizing L_{FACD}, the student network learns to place positive pairs closer and negative pairs further apart. Through this approach, mutual information between the feature maps of the teacher and student networks can be maximized [32]. The effectiveness of contrastive loss on each domain is described in our ablation studies.

C. OVERALL LOSS FUNCTION

The overall loss function of FACD is constructed via contrastive distillation in the image and feature domains, which is formulated as follows:

$$L_{total} = L_{SR} + \lambda L_{FACD} \tag{5}$$

where λ denotes a hyperparameter for balancing L_{SR} and L_{FACD}. The hyperparameter of λ was set to 4 in our experimental configurations.

IV. EXPERIMENTS

In this section, we explain the details of our experimental network configurations and analyze the results both quantitatively and qualitatively.

A. EXPERIMENTAL CONFIGURATIONS

Following previous works [20], [21], [22], [25], [36], [39], we used 800 split set images from the DIV2K dataset [57] for training. FACD was also evaluated with luminance-peak-signal-to-noise-ratio (Y-PSNR) on four benchmark datasets—Set 5 [58], Set 14 [59], BSD 100 [60], and Urban 100 [61]. For comparison with previous KD algorithms, we performed experiments on existing SISR networks, EDSR [22] and RCAN [36]. Table 1 lists the configuration details of distillation models consisting of teacher and student networks. The configuration of each distillation model was identical as in the respective previous works to ensure fair experimental comparison. While EDSR reduces the number of residual blocks (ResBlocks) and the channel size of the convolution, RCAN retains the number of residual groups (ResGroups) containing multiple ResBlocks, and only reduces the number of ResBlocks. This distillation compresses EDSR by approximately a factor of 30, and RCAN by approximately a factor of 3 in terms of the number of parameters. In details, each FD used the same configuration of teachers and students, so the computational and memory complexity of each FD scheme is the same.

TABLE 1. Network descriptions of teacher and student networks. T and S denote the teacher and student network respectively.

	EDSR		RCAN	
	T	S	T	S
Channel size	256	64	64	64
ResBlocks	32	16	20	6
ResGroups	-	-	10	10
Params (M)	43	1.5	15.59	5.17

FACD was implemented using PyTorch 1.8.0 with an NVIDIA TITAN RTX GPU. All student networks using distillation were trained using the ADAM optimizer with default hyper-parameters in PyTorch. Unlike the configurations used in previous works (200 [20] or 300 [21] epochs), FACD loss was not sufficiently saturated. Therefore, the batch size and total number of epochs were set to 16 (same as in the previous works) and 600, respectively. The initial learning rate was set to 2×10^{-4}, and was halved after 150 epochs. In addition, the patch size for training was set to 48×48 for the network input, and the default configurations were applied for data augmentation (e.g., horizontal flip, vertical flip, and random rotation). The same experimental configurations were applied to FAKD [20], LFD [21], CSD [25], and PISR [48], which are the main comparison works in this paper, to ensure reproducible results.

B. QUANTITATIVE RESULTS

The quantitative results in terms of Y-PSNR are presented in Table 2. FACD achieved the best performance on almost all benchmark datasets and scale factors, except for the Set 5 dataset on EDSR x4. Compared to the conventional FD, both EDSR and RCAN exhibited an average performance improvement of approximately 0.07 dB. The performance improvement in RCAN was greater than that in EDSR. The most significant distinction between EDSR and RCAN was the inclusion of feature-attention blocks. As a result, rendering the features more similar to RCAN using the feature attention scheme was effective. In other words, the knowledge of teacher networks can be better utilized in RCAN than in EDSR in the feature domain.

1) IMPACT ON SCALE FACTOR

Table 3 summarizes the evaluation results presented in Table 2. PSNR performance improvement is the average difference between the performance of FACD and the overall FD performance, and is listed in Table 2. As presented in Table 3, the degree of performance improvement in FACD, compared to other FD methods, decreased as the scale factor increased. In general, the performance improvement efficiency of scale x2 was approximately two times better than that of scale x4. As the scale factor increased, texture restoration became more difficult, which induced an upper bound on the performance of the teacher network. This implies that the knowledge that can be transmitted by the teacher network is limited at larger scale factors.

2) PERFORMANCE ON URBAN100

Each benchmark dataset for SISR exhibits its own data characteristics. For instance, Sets 5 and 14 contain samples of simple objects, and BSD100 exhibits various characteristics, ranging from natural images to complex textures. The Urban100 dataset includes a variety of repeated patterns and edges that arise from the complex architecture of buildings. As presented in Tables 2 and 3, FACD exhibited better quantitative performance, especially on the Urban100 dataset. Thus, FACD has an advantage over other FD approaches in terms of texture restoration. FACD achieved a PSNR improvement of 0.56 dB, 0.34 dB, 0.17 dB, and 0.09dB over the baseline student, FAKD, LSFD, and PISR on scale x2 in the RCAN network, respectively.

3) COMPARISON WITH CSD

Consequently, CSD of the teacher and student networks has an identical number of ResBlocks, except for the number of channels. However, as depicted in Fig. 1, the teacher and student networks of EDSR exhibit different numbers of ResBlocks and channels. RCAN teacher and student networks include different numbers of ResBlocks and ResGroups; however, they exhibits the same number of channels.

Therefore, to ensure a fair experimental comparison of the distillation domain with contrastive loss, we configured the teacher model identically to CSD (R16C256) and compared their distillation performances in the EDSR networks. Furthermore, CSD was compared with FCD to exclude its effect on adaptive distillation. As presented in Table 4, FCD exhibited an average PSNR improvement of 0.02 dB compared to CSD. Thus, the domain applying the contrastive distillation was more efficient on the features in the network than on the feature of the output.

4) COMPARISON WITH PRUNING METHODS

To further demonstrate the effectiveness of the proposed FD method, we compared the evaluation results with SOTA pruning methods such as ASSL [26] and SRP [27]. To compare each methods in terms of complexity, Floating point Operations (FLOPs) measure when the output image size is set to $3 \times 1280 \times 720$ in the inference process. To ensure fair comparison, the student model was configured for KD to match the computational cost of each pruned EDSR sub-network model as closely as possible. As presented in Table 5, FACD achieved an average PSNR improvement of 0.02 dB over the SOTA pruning methods at smaller network sizes.

C. QUALITATIVE RESULTS

As depicted in Fig. 4 and 5, we qualitatively compared FACD with existing approaches on the Set14/Urban100 benchmark datasets. To compare the difference in the restoration quality of detailed patterns, we performed the comparison on relatively small cropped images. PSNR scores were calculated with respect to only the cropped images.

TABLE 2. Quantitative results (PSNR) measured by applying different FD methods on the student EDSR and RCAN network, as shown in Table 1. Note that the model efficiency such as running time and memory for each FD method is the same. Note that FAKD*, LFD* and PISR* indicate our reproduced results with our experimental settings. Except for them, the results in the table are taken from their respective paper. **Red** indicates the best PSNR within each dataset, and **Blue** indicates the second best.

EDSR						RCAN					
Methods	Scale	Set5	Set14	B100	Urban100	Methods	Scale	Set5	Set14	B100	Urban100
Teacher	x2	38.190	33.857	32.351	32.873	Teacher	x2	38.271	34.126	32.390	33.176
Student		37.919	33.439	32.102	31.728	Student		38.074	33.623	32.199	32.317
FAKD*		37.976	33.523	32.156	31.906	FAKD*		38.164	33.815	32.274	32.533
LFD		37.984	33.547	32.156	31.896	LFD		38.178	33.840	32.296	32.669
LFD*	x2	37.986	33.528	32.159	31.935	LFD*	x2	38.180	33.851	32.305	32.681
LSFD		37.991	33.529	32.157	31.936	LSFD		38.189	33.882	32.323	32.704
PISR*		37.971	33.557	32.166	32.025	PISR*		38.254	33.941	32.326	32.789
FACD (ours)		38.043	33.588	32.188	32.072	FACD (ours)		38.242	34.016	32.334	32.878
Teacher	x3	34.547	30.435	29.167	28.470	Teacher	x3	34.758	30.627	29.309	29.104
Student		34.272	30.266	29.044	27.959	Student		34.557	30.408	29.162	28.482
FAKD*		34.356	30.296	29.066	28.016	FAKD*		34.653	30.449	29.208	28.523
LFD		34.348	30.287	29.068	27.999	LFD		34.657	30.525	29.224	28.665
LFD*	x3	34.333	30.301	29.077	28.022	LFD*	x3	34.659	30.515	29.226	28.672
LSFD		34.384	30.302	29.077	28.029	LSFD		34.666	30.510	29.227	28.689
PISR*		34.362	30.317	29.083	28.058	PISR*		34.691	30.533	29.215	28.751
FACD (ours)		34.394	30.333	29.103	28.125	FACD (ours)		34.729	30.563	29.262	28.818
Teacher	x4	32.385	28.741	27.661	26.425	Teacher	x4	32.638	28.851	27.748	26.748
Student		32.102	28.526	27.538	25.905	Student		32.321	28.688	27.634	26.340
FAKD*		32.138	28.547	27.557	25.972	FAKD*		32.461	28.750	27.678	26.422
LFD		32.107	28.524	27.552	25.962	LFD		32.475	28.783	27.693	26.542
LFD*	x4	32.099	28.557	27.546	25.962	LFD*	x4	32.479	28.774	27.688	26.547
LSFD		32.107	28.548	27.563	25.980	LSFD		32.497	28.711	27.699	26.525
PISR*		32.110	28.573	27.559	25.987	PISR*		32.512	28.796	27.689	26.538
FACD (ours)		32.128	28.580	27.580	26.029	FACD (ours)		32.540	28.810	27.708	26.606

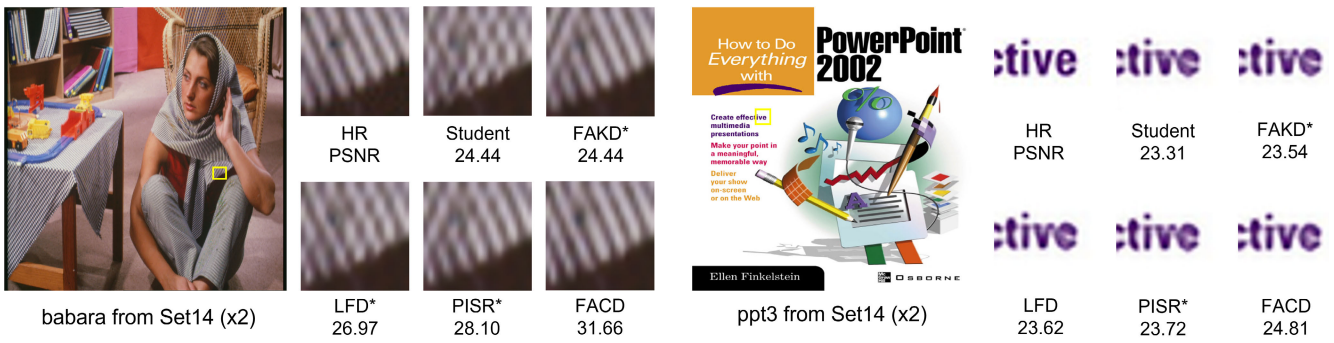

FIGURE 4. Qualitative results on EDSR with scale x2. Noted that FAKD* and LFD* indicate our reproduced results with the same experimental settings.

TABLE 3. Evaluation results on average PSNR improvement over other FD approaches. Performance improvements that are greater than 0.1dB are marked with an underline.

Model	scale	Set 5	Set 14	B100	Urban100
EDSR	x2	+0.063	+0.053	+0.030	+0.132
	x3	+0.037	+0.029	+0.029	+0.100
	x4	+0.016	+0.030	+0.025	+0.056
RCAN	x2	+0.049	+0.150	+0.029	+0.203
	x3	+0.063	+0.057	+0.042	+0.158
	x4	+0.055	+0.047	+0.019	+0.091

1) QUALITATIVE RESULTS COMPARED TO OTHER FD METHODS

In general, as evidenced by the qualitative results, PSNR performance is proportional to the subjective image quality. Our results clearly confirmed that FACD achieved better PSNR and qualitative results than other FD approaches. In particular, in terms of texture restoration (e.g. patterns), FACD yielded clearer textures and exhibited greater similarity between the teacher and HR images than other FD as shown in Figs. 4 and 5. In the case of scale x4 images (078 from Urban 100), a difference was confirmed in the distortion of texture as shown in Fig. 5. In other cases (babara/ppt3 from Set14, x2), a difference was observed in the sharpness and straightness of the line as shown in Fig. 4.

2) QUALITATIVE RESULTS ON RCAN NETWORK

Now, we present qualitative results of FD on the RCAN network. As depicted in Fig. 6, FACD achieved better PSNR scores and qualitative results compared with other FD approaches. In particular, FACD recovered the linear patterns presented in Fig. 6 better than conventional FD.

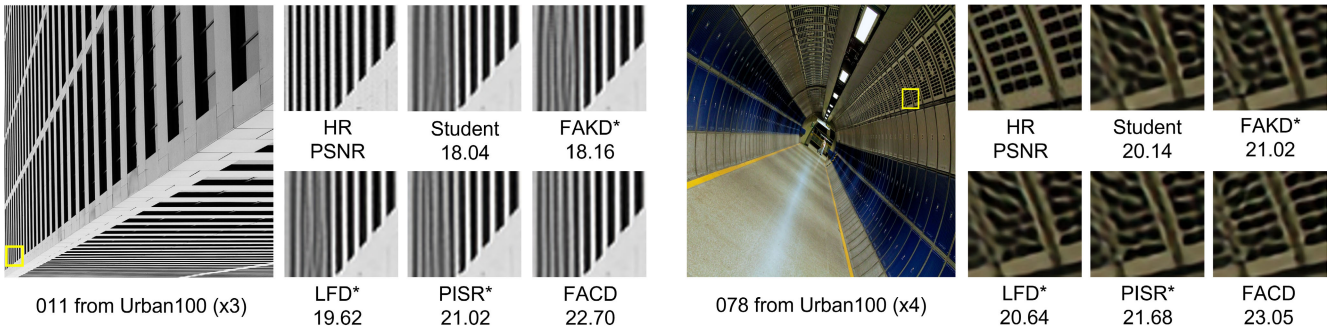

FIGURE 5. Qualitative results on EDSR with scale x3 and x4. Noted that FAKD* and LFD* indicate our reproduced results with the same experimental settings.

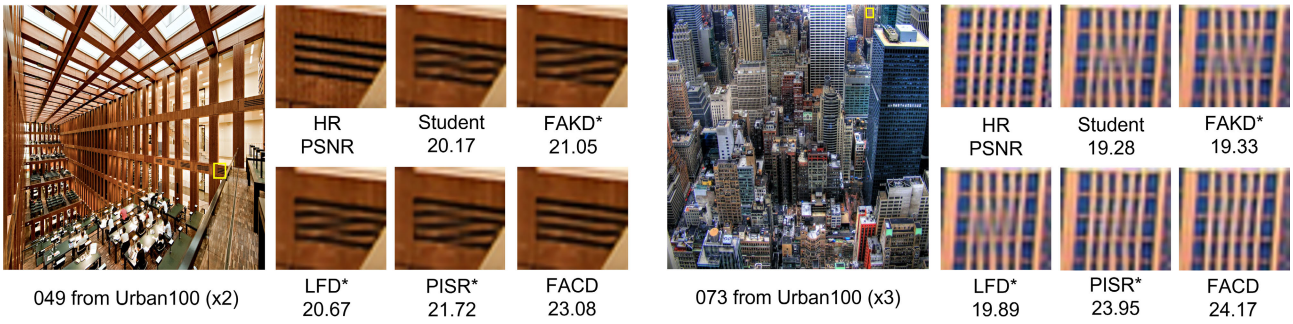

FIGURE 6. Qualitative results on RCAN with scale factors of x2 and x3 super-resolution (SR). Note that feature affinity-based knowledge distillation (FAKD*) and local feature distillation (LFD*) represent the results reproduced using identical experimental configurations.

TABLE 4. Quantitative results (PSNR) of CSD and FCD. Note that CSD* indicates our reproduces results with our experimental settings. Both networks of teacher configuration are same. Better result is marked in Red.

Methods	scale	Set 5	Set 14	B100	Urban100
CSD*	x2	38.001	33.536	32.160	31.984
FCD	x2	38.015	33.552	32.174	32.023
CSD*	x3	34.378	30.309	29.082	28.020
FCD	x3	34.379	30.318	29.099	28.086
CSD*	x4	32.112	28.563	27.569	25.988
FCD	x4	32.121	28.578	27.571	26.011

V. ABLATION STUDY

This section demonstrates the effectiveness of FACD and presents an ablation study conducted to evaluate the effectiveness of each loss component, formulation of contrastive loss, the effects of adaptive distillation in the image and feature domains, that of contrastive distillation in the feature and spatial affinity (SA) matrix domains, and the fidelity of the distillation schemes.

1) IMPACT ON CONTRASTIVE LOSS DOMAINS

To demonstrate the effectiveness of contrastive loss in the feature domain, we compared the contrastive loss results in the feature and image domains. The composition of contrastive loss, including the formation of the equation, was identical in the two cases, except for the application domain. As presented in Table 6, FCD outperformed both ICD and CSD. Feature-domain contrastive distillation achieved a PSNR improvement of 0.04 dB compared to image-domain contrastive distillation.

To demonstrate the effect of applying a contrastive loss for each domain, the qualitative results of CSD and ICD were analyzed. As depicted in Fig. 6, FCD achieved better PSNR scores and qualitative results than other CD approaches. FCD restored both straight lines depicted in Fig. 6, while the student network and FAKD restored only one. Feature domain with contrastive loss performed better than the image domain in terms of texture restoration.

2) IMPACT ON CONTRASTIVE LOSS FORMULATION

InfoNCE loss has been primarily used in contrastive learning and unsupervised learning [63]. Similarity measures of contrastive loss in InfoNCE loss use the dot-product operation. On the other hand, FACD uses Euclidean distance-based loss as a similarity measure for contrastive loss. As presented in Table 7, contrastive loss based on Euclidean distance improved PSNR by a margin of 0.02 dB compared to InfoNCE loss.

3) IMPACT ON EACH LOSS COMPONENT

To confirm the effect on the performance of each loss component, each loss component was turned on/off and tested. L1_GT denotes the conventional Euclidean loss over GT images, and L1_T denotes the image domain distillation loss. The difference between FCD and FACD is whether or not adaptive distillation is used. The overall results are presented in Table 8. Compared with the baseline model without distillation, the proposed distillation approach achieved significant performance improvement on all benchmark datasets.

TABLE 5. PSNR (dB) results on the Set5 (x2) in the EDSR sub-network. Note that ASSL [26] and SRP [27] indicate the original paper results. (Red indicates best PSNR within the same network settings).

Params (K)		FLOPs (G)		Pruning				KD
Pruning	Ours	Pruning	Ours	Scratch	L1-norm [62]	ASSL [26]	SRP [27]	FACD (ours)
1101.8	1087.2	254.5	251.1	37.85	37.91	37.94	37.97	37.99
681.1	678.5	157.5	156.9	37.81	37.81	37.91	37.89	37.93
381.8	378.8	88.9	88.2	37.75	37.73	37.82	37.84	37.86
154.2	153.8	36.5	36.2	37.56	37.58	37.70	37.71	37.73
26.9	26.2	7.3	7.1	36.74	36.87	37.23	37.28	37.31

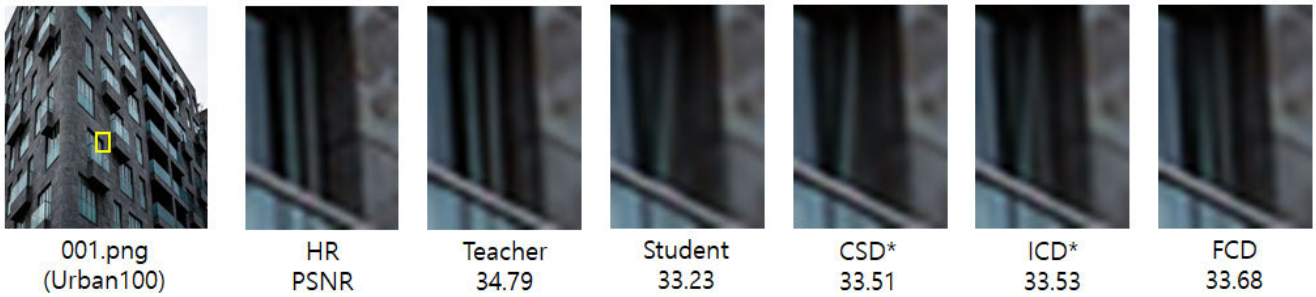

FIGURE 7. Qualitative results on residual channel attention networks (RCAN) with a scale factor of x2 super-resolution (SR). Note that contrastive self-distillation (CSD*) and image-domain contrastive distillation (ICD*) represent the results reproduced using identical experimental configurations.

TABLE 6. Ablation study (PSNR) on contrastive loss comparison between image and feature domains (x2, EDSR). FCD refers to the method obtained by removing the adaptive scheme from FACD. CSD* represents the results reproduced using our experimental configurations.

Methods	Distillation Domain	Set5	Set14	B100	Urban100
ICD	Image	37.983	33.558	32.159	32.020
CSD* [25]	VGG features of image	38.001	33.536	32.160	31.944
FCD(ours)	Intermediate features	38.025	33.581	32.183	32.064

TABLE 7. Ablation study (PSNR) of the different contrastive loss (L_{FACD}) on EDSR x2.

Methods	Set5	Set14	B100	Urban100
FCD(InfoNCE)	38.005	33.556	32.173	32.028
FCD(ours)	38.025	33.581	32.183	32.064

The 4-th row in Table 8 demonstrates that training was efficient when distillation was applied independently without L1_GT. In particular, the FCD or FACD components exhibited larger performance improvements compared with other loss components, as evidenced by the performance difference shown between the 1-st and 3-rd rows. This implies that the proposed FCD or FACD transferred the knowledge from the teacher network to the student network effectively. Finally, the combination of all loss components achieved the best evaluation results on various benchmark datasets.

4) IMPACT ON ADAPTIVE DISTILLATION

In Section III-A, we describe the impact of worse cases obtained from the teacher networks. To show the effectiveness of adaptive distillation scheme, we compared the quantitative results obtained using the proposed FACD and FCD (FACD without an adaptive distillation approach). As presented in Table 9, FACD achieved a PSNR improvement of

TABLE 8. Ablation study (PSNR) on the effectiveness of each loss component in the RCAN network (x4). Red indicates the best PSNR within each dataset.

Loss Component				Set5	Set14	B100	Urban100
L1_GT	L1_T	FCD	FACD				
✓				32.321	28.688	27.634	26.340
✓	✓			32.362	28.722	27.657	26.402
✓		✓		32.425	28.738	27.669	26.507
	✓	✓		32.447	28.768	27.671	26.524
✓	✓	✓		32.492	28.774	27.689	26.562
✓	✓		✓	32.540	28.810	27.708	26.606

0.02 dB compared to FCD. This ablation study confirmed the importance of the adaptive distillation approach.

5) IMPACT ON FEATURE ATTENTION

The three intermediate feature matching points were configured for fair FD comparison. To compare the effects of the different features, we compared the evaluation results of attention corresponding to each feature point. The only difference among the three methods was the composition of the w_j in Eq. 4. As presented in Table 10, FCD with the proposed attention version (FAT) achieved the best performance in terms of PSNR. This indicates that, due to the cascading architecture of the CNN-based SR network, the upper part of the network wields greater influence on the distillation performance than the lower part.

6) EFFECT OF ADAPTIVE DISTILLATION IN THE FEATURE DOMAIN

In this section, we describe the effect of applying adaptive distillation to the feature domain. Because feature information is more important than image information owing to the cascading architecture of the CNN-based SR network,

TABLE 9. Ablation study (PSNR) on the effectiveness of adaptive distillation methods in the EDSR and RCAN networks. FCD indicates feature-domain contrastive distillation without adaptive distillation. Red indicates the best PSNR within each dataset.

Methods	Scale	Model			
		EDSR		RCAN	
		B100	Urban100	B100	Urban100
FCD	x2	32.183	32.064	32.331	32.851
FACD		32.188	32.072	32.334	32.878
FCD	x3	29.096	28.124	29.242	28.771
FACD		29.103	28.125	29.262	28.818
FCD	x4	27.578	26.013	27.689	26.562
FACD		27.580	26.029	27.708	26.606

TABLE 10. Ablation study (PSNR) of the feature attention on EDSR x2. FAA, FAB, and FAT indicate paying attention to the average, bottom, and top parts of features matching points, respectively. FAU is the version of this paper. The better performance is marked in Red.

Methods	$[w_1, w_2, w_3]$	Set5	Set14	B100	Urban100
FCD(FAA)	[0.33, 0.33, 0.33]	38.007	33.579	32.177	32.058
FCD(FAB)	[0.20, 0.30, 0.50]	37.998	33.577	32.175	32.050
FCD(FAT)	[0.50, 0.30, 0.20]	38.025	33.581	32.183	32.064

we focused on the effect of adaptive distillation in the feature domain.

The composition of feature domain loss with an adaptive scheme is formulated as follows:

$$L_{FACD}^{on} = \sum_{i=1}^{N}\sum_{j=1}^{3} w_j \frac{\alpha_i \| DR(\hat{F}_{ij}^{S}) - \hat{F}_{ij}^{T} \|_1}{\sum_{k=1}^{K} \| DR(\hat{F}_{kj}^{S}) - \hat{F}_{kj}^{Neg} \|_1} \tag{6}$$

where α_i denotes the indicator of inappropriate teacher samples, and descriptions of the formulation are identical to those in the Section III.

In other words, the composition of feature loss without the adaptive scheme is formulated as follows:

$$L_{FACD}^{off} = \sum_{i=1}^{N}\sum_{j=1}^{3} w_j \frac{\| DR(\hat{F}_{ij}^{S}) - \hat{F}_{ij}^{T} \|_1}{\sum_{k=1}^{K} \| DR(\hat{F}_{kj}^{S}) - \hat{F}_{kj}^{Neg} \|_1} \tag{7}$$

where the descriptions of the formulation are also identical to those in Eq. 4, except for α_i.

As presented in Table 11, FD with adaptive distillation outperformed adaptive distillation on the image domain only.

7) IMPACT OF LOSS TYPE ON EACH DOMAIN

To confirm the effect of contrastive distillation on each domain, we perform an experimental comparison by properly distinguishing the loss types (e.g., Euclidean distance and contrastive loss) and the domains of application (e.g., SA and feature). The spatial affinity (SA) matrix in FAKD represents the spatial correlation between pixels [20] and is formulated as follows:

$$SA = \hat{F}^T \times \hat{F} \tag{8}$$

where $\hat{F}$ denotes the normalized feature map. The dimensions of SA are $HW \times HW$, where H and W denote the height and width of the input image, respectively.

TABLE 11. Quantative results on the effectiveness of adaptive distillation in the feature domain (EDSR, x4). The formulation of L_{SR} is described as the Section III.

L_{total}	Set5	Set14	B100	Urban100
$L_{SR} + \lambda L_{FACD}^{off}$	32.120	28.571	27.568	26.013
$L_{SR} + \lambda L_{FACD}^{on}$	32.129	28.580	27.580	26.029

TABLE 12. Quantative results on the effectiveness of contrastive loss in each domain (EDSR, x2). CD is an abbreviation for Contrastive Distillation. Noted that SA indicates spatial affinity matrix which is used in the FAKD paper. The best performance is marked in Red.

Methods	Descriptions	Set5	Set14	B100	Urban100
FAKD*	L1 in SA	37.976	33.523	32.156	31.906
LFD*	L1 in Feat.	37.986	33.518	32.159	31.935
FCD(SA)	CD in SA	37.992	33.571	32.180	32.045
FCD(ours)	CD in Feat.	38.025	33.581	32.183	32.064

TABLE 13. Evaluation results of the average PSNR between the output images of student and teacher in RCAN networks. FAKD* and LFD* indicate our reproduced results with our experimental settings. The best performance in the same setting is marked in Red.

Methods	scale	Set5	Set14	B100	Urban100
FAKD*	x2	49.581	43.808	45.342	38.917
LFD*		50.120	44.231	46.633	39.351
PISR*		50.623	44.828	46.823	40.551
FACD (ours)		50.882	45.019	46.881	40.757
FAKD*	x3	45.390	40.677	42.472	35.279
LFD*		45.518	41.919	42.883	35.433
PISR*		45.887	42.007	43.135	36.551
FACD (ours)		45.922	42.083	43.397	36.684
FAKD*	x4	42.137	38.562	40.461	33.210
LFD*		42.233	39.019	40.888	33.677
PISR*		42.820	39.231	41.106	33.951
FACD (ours)		42.954	39.481	41.358	34.088

To ensure fair comparison, the adaptive distillation scheme was removed, and the performance of the resulting architecture was evaluated. As presented in Table 12, the proposed distillation scheme achieved significant enhancement in terms of PSNR on all benchmark datasets. We also confirmed that SA leads to learning feature correlations from each pair of pixels, even if it comprises unrelated pairs (e.g., different objects). For this reason, FD is more effective than SA in feature domains with a deep regressor [21]. Moreover, as described in the Section. IV, contrastive loss performed better than L1-distance loss in the FD approach.

8) PERFORMANCE OF THE FD SCHEME OVER THE TEACHER NETWORKS

To compare the similarity between each FD and the teacher networks, we evaluated the PSNR between the output images of teacher and student networks. During this process, the output images of the teacher were considered as the target (GT). FACD was exhibited the best performance in all test configurations. As presented in Table 13, FACD exhibited an average PSNR improvement of 0.60 dB compared to other FD methods. This indicates that feature-based contrastive loss improves the similarity between output images of student and teacher networks.

VI. CONCLUSION

In this paper, we analyzed the limitation of the conventional FD scheme with Euclidean distance-based loss and the

impact of inappropriate results of the teacher network on the distillation performance, and proposed the FACD that employs adaptive contrastive distillation in the feature domain. In detail, to transfer knowledge from the teacher networks more effectively, we proposed the FD based on contrastive loss, which maximizes the mutual information of features between the student and teacher networks, and the adaptive distillation scheme by rejecting inaccurate results of knowledge transfer from the teacher network during the KD process. The student networks of EDSR and RCAN based on FACD achieved SOTA results in KD for SISR, and produced the excellent qualitative results in terms of texture reconstruction. Since FACD is a simple method that effectively enhanced the performance using teacher models, we plan to extend the FACD to other low-level vision networks (e.g., restoration, real-world SR, etc.) that operate in real-time on industrial sites (e.g., smartphone devices) [28], [64] and other types of networks such as transformer. However, the issues of training cost (e.g., 600 epochs for saturation) and semantic collapse due to instance-wise contrastive loss remain to be addressed.

REFERENCES

[1] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 25–47, 2000.

[2] D. Qiu, L. Zheng, J. Zhu, and D. Huang, "Multiple improved residual networks for medical image super-resolution," *Future Gener. Comput. Syst.*, vol. 116, pp. 200–208, Mar. 2021.

[3] Y. Xiao, Q. Yuan, K. Jiang, X. Jin, J. He, L. Zhang, and C.-W. Lin, "Local-global temporal difference learning for satellite video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Sep. 5, 2023, doi: 10.1109/TCSVT.2023.3312321.

[4] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610819.

[5] Y. Xiao, Q. Yuan, Q. Zhang, and L. Zhang, "Deep blind super-resolution for satellite video," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5516316.

[6] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.

[7] K. Jiang, Z. Wang, P. Yi, and J. Jiang, "Hierarchical dense recursive network for image super-resolution," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107475.

[8] K. Jiang, Z. Wang, P. Yi, J. Jiang, J. Xiao, and Y. Yao, "Deep distillation recursive network for remote sensing imagery super-resolution," *Remote Sens.*, vol. 10, no. 11, p. 1700, Oct. 2018. [Online]. Available: https://www.mdpi.com/2072-4292/10/11/1700

[9] J. Jiang, C. Wang, X. Liu, K. Jiang, and J. Ma, "From less to more: Spectral splitting and aggregation network for hyperspectral face super-resolution," 2021, *arXiv:2108.13584*.

[10] K. Jiang, Z. Wang, P. Yi, T. Lu, J. Jiang, and Z. Xiong, "Dual-path deep fusion network for face image hallucination," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 378–391, Jan. 2022.

[11] K. Jiang, Z. Wang, P. Yi, G. Wang, K. Gu, and J. Jiang, "ATMFN: Adaptive-threshold-based multi-model fusion network for compressed face hallucination," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2734–2747, Oct. 2020.

[12] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 742–751.

[13] K. Hayat, "Multimedia super-resolution via deep learning: A survey," *Digit. Signal Process.*, vol. 81, pp. 198–217, Oct. 2018.

[14] Z. Shao, L. Wang, Z. Wang, and J. Deng, "Remote sensing image super-resolution using sparse representation and coupled sparse autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2663–2674, Aug. 2019.

[15] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[16] F. Zhou, W. Yang, and Q. Liao, "Interpolation-based image super-resolution using multisurface fitting," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3312–3318, Jul. 2012.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[18] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[19] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[20] Z. He, T. Dai, J. Lu, Y. Jiang, and S.-T. Xia, "Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 518–522.

[21] S. Park and N. Kwak, "Local-selective feature distillation for single image super-resolution," 2021, *arXiv:2111.10988*.

[22] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[23] Z. Hou and S.-Y. Kung, "Efficient image super resolution via channel discriminative deep neural network pruning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3647–3651.

[24] Z. Luo, Y. Li, L. Yu, Q. Wu, Z. Wen, H. Fan, and S. Liu, "Fast nearest convolution for real-time efficient image super-resolution," 2022, *arXiv:2208.11609*.

[25] Y. Wang, S. Lin, Y. Qu, H. Wu, Z. Zhang, Y. Xie, and A. Yao, "Towards compact single image super-resolution via contrastive self-distillation," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1122–1128, doi: 10.24963/ijcai.2021/155.

[26] Y. Zhang, H. Wang, C. Qin, and Y. Fu, "Aligned structured sparsity learning for efficient image super-resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2695–2706.

[27] y. Zhang, H. Wang, C. Qin, and Y. Fu, "Learning efficient image super-resolution networks via structure-regularized pruning," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–12.

[28] G. Gankhuyag, J. Huh, M. Kim, K. Yoon, H. Moon, S. Lee, J. Jeong, S. Kim, and Y. Choe, "Skip-concatenated image super-resolution network for mobile devices," *IEEE Access*, vol. 11, pp. 4972–4982, 2023.

[29] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[30] N. Aghli and E. Ribeiro, "Combining weight pruning and knowledge distillation for CNN compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3185–3192.

[31] J. Kim, S. Chang, and N. Kwak, "PQK: Model compression via pruning, quantization, and knowledge distillation," 2021, *arXiv:2106.14681*.

[32] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," 2019, *arXiv:1910.10699*.

[33] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, and P. Luo, "DetCo: Unsupervised contrastive learning for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8372–8381.

[34] A. Romero, N. Ballas, S. Ebrahimi Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.

[35] H. Moon, Y.-H. Kwon, J. Jeong, and S. Kim, "Compression of super-resolution model using contrastive learning," in *Proc. Korean Inst. Broadcast Media Eng. Conf.*, 2022, pp. 557–559.

[36] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 286–301.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[38] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[39] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11057–11066.
[40] Z. Du, D. Liu, J. Liu, J. Tang, G. Wu, and L. Fu, "Fast and memory-efficient network towards efficient image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 852–861.
[41] Y. Mao, N. Zhang, Q. Wang, B. Bai, W. Bai, H. Fang, P. Liu, M. Li, and S. Yan, "Multi-level dispersion residual network for efficient image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 1660–1669.
[42] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.
[43] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 63–79.
[44] K. C. K. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "GLEAN: Generative latent bank for large-factor image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14240–14249.
[45] K. C. K. Chan, X. Xu, X. Wang, J. Gu, and C. C. Loy, "GLEAN: Generative latent bank for image super-resolution and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3154–3168, Mar. 2023.
[46] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*.
[47] M. R. U. Saputra, P. Gusmao, Y. Almalioglu, A. Markham, and N. Trigoni, "Distilling knowledge from a deep pose regressor network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 263–272.
[48] W. Lee, J. Lee, D. Kim, and B. Ham, "Learning with privileged information for efficient image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 465–482.
[49] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.
[50] M. Kim, J. Tack, and S. J. Hwang, "Adversarial self-supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 2983–2994.
[51] J. Zhu, S. Tang, D. Chen, S. Yu, Y. Liu, M. Rong, A. Yang, and X. Wang, "Complementary relation contrastive distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9256–9265.
[52] G. Wu, J. Jiang, and X. Liu, "A practical contrastive learning framework for single-image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 10, 2023, doi: 10.1109/TNNLS.2023.3290038.
[53] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution," *Inf. Fusion*, vol. 96, pp. 297–311, Aug. 2023.
[54] J. Zhang, S. Lu, F. Zhan, and Y. Yu, "Blind image super-resolution via contrastive representation learning," 2021, *arXiv:2107.00708*.
[55] D. Mishra and O. Hadar, "CLSR: Contrastive learning for semi-supervised remote sensing image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[57] R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1110–1121.
[58] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L.-A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–32.
[59] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.* Cham, Switzerland: Springer, 2010, pp. 711–730.
[60] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, Jul. 2001, pp. 416–423.
[61] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.
[62] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient ConvNets," 2016, *arXiv:1608.08710*.
[63] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[64] G. Gankhuyag, K. Yoon, J. Park, H. Seon Son, and K. Min, "Lightweight real-time image super-resolution network for 4K images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 1746–1755.

HYEON-CHEOL MOON received the B.S. and M.S. degrees in electronics and information engineering from Korea Aerospace University, Goyang, Republic of Korea, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. Since 2021, he has been a Research Engineer with the Intelligent Image Processing Research Center, Korea Electronics Technology Institute, Seongnam, South Korea. His research interests include multimedia signal processing and artificial intelligence processing.

JAE-GON KIM (Member, IEEE) received the B.S. degree in electronics engineering from Kyungpook National University, Daegu, South Korea, in 1990, and the M.S. and Ph.D. degrees in electronic engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1992 and 2005, respectively. From 1992 to 2007, he was with the Electronics and Telecommunications Research Institute (ETRI), where he has involved in the development of digital broadcasting media services, the MPEG-2/4/7/21 standards and related applications, and convergence media technologies. From 2001 to 2002, he was a Staff Associate with the Department of Electrical Engineering, Columbia University, New York, USA. Since 2007, he has been with Korea Aerospace University, Goyang, South Korea, where he is currently a Professor with the School of Electronics and Information Engineering. From 2014 to 2015, he was a Visiting Scholar with the Video Signal Processing Laboratory, University of California San Diego. He has involved in video coding standards in JCT-VC and JVET activities of ITU-T VCEG and ISO/IEC MPEG. His research interests include image/video compression, video signaling processing, immersive video, and artificial intelligence processing.

JINWOO JEONG received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2004, 2006, and 2011, respectively. From 2011 to 2015, he was a Senior Video Signal Processing Engineer with Samsung Electronics Company Ltd., Suwon, South Korea. Since 2016, he has been a Principal Research Engineer with the Intelligent Image Processing Research Center, Korea Electronics Technology Institute, Seongnam, South Korea. His research interests include multimedia signal processing, artificial intelligence processing, and VR/AR technologies.

SUNGJEI KIM received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2004, 2006, and 2011, respectively. From 2011 to 2015, he was a Senior Video Signal Processing Engineer with Samsung Electronics Company Ltd., Suwon, South Korea. Since 2015, he has been a Principal Research Engineer with the Intelligent Image Processing Research Center, Korea Electronics Technology Institute, Seongnam, South Korea. His research interests include multimedia signal processing, artificial intelligence processing, and VR/AR technologies.