Multi-Scale Contrastive-Adversarial Distillation for Super-Resolution

Donggeun Ko[1], Youngsang Kwak[1], San Kim[2], Jaehwa Kwak[1], Jaekwang Kim[2]
[1]AiM Future Inc. [2]Sungkyunkwan University
{sean.ko, youngsang.kwak, jaehwa.kwak}@aimfuture.ai, {saankim, linux}@skku.edu

Abstract

Knowledge distillation (KD) is a powerful technique for model compression, enabling the creation of compact and efficient "student" models by transferring knowledge from large-scale, pre-trained "teacher" models. However, the application of traditional KD methods in this domain is considerably more challenging than in high-level tasks like classification, as the SISR task is to reconstruct image pixels a regression problem. Hence, to effectively distill the knowledge of a teacher model in SR, we propose MCAD-KD, Multi-Scale Contrastive–Adversarial Distillation for Super-Resolution. We utilize a novel hybrid contrastive learning framework that operates on both global (image-level) and local (patch-level) scales. Furthermore, we integrate adversarial guidance, which pushes the student's output towards the manifold of realistic images, allowing it to potentially surpass the perceptual quality of the teacher by learning directly from the ground-truth data distribution. Our comprehensive framework synergistically combines these components to train a lightweight student model that achieves a superior trade-off between perceptual quality and computational efficiency.

1. Introduction

Image super-resolution (SR) is a foundational task in computer vision (CV) that seeks to reconstruct a high-resolution (HR) image from a low-resolution (LR) input, with applications ranging from medical imaging to surveillance and consumer electronics. Early deep-learning approaches—such as SRCNN [8], RCAN [35] and EDSR [21]—demonstrated that convolutional neural networks (CNNs) can achieve impressive PSNR and SSIM gains. More recently, Transformer-based architectures like SwinIR [20], and IPT [6] have further advanced perceptual quality by modeling long-range dependencies.

Despite these advances, both CNN- and Transformer-based SR models demand extensive computational resources and memory, limiting their deployment on edge devices and in real-world scenarios. Quantization [11], and pruning [2] alleviate some cost but often require specialized hardware or extensive tuning. Knowledge distillation (KD) offers a complementary strategy by transferring "dark knowledge" from a large teacher network to a compact student network [14]. KD has proven highly effective in natural language processing [30] and high-level CV tasks such as classification [29], detection [5], and segmentation [23]. However, vanilla KD methods that match soft logits or intermediate features often yield marginal gains, or even degrade performance, when applied to SR networks. In SR, the teacher's output is itself an approximation of the ground truth, providing little extra information for pixel-wise or feature-wise imitation. Recent SR-specific distillation methods exploit data upcycling and label consistency regularization [36], but still rely on pixel-level or feature-level alignment that can propagate teacher artifacts. To address these limitations, we propose MCAD-KD, a unified SR distillation framework that combines multi-scale contrastive distillation and adversarial distillation guidance. Multi-scale contrastive objectives align global structure and local texture relations in a shared embedding space, avoiding direct pixel mimicking. Adverarial distillation guidance provides discriminative feedback against real HR images with a discriminator, D, driving the student to produce perceptually superior outputs beyond the teacher's capabilities. Thus, our contributions are summarized as follows: 1) A novel multi-scale contrastive distillation strategy that transfers relational knowledge at both image and patch levels. 2) Integration of adversarial distillation guidance into KD, enabling the student to surpass teacher performance in perceptual quality. 3) Comprehensive experiments demonstrate that our MCAD-KD outperforms other baselines.

2. Related Works

Image Super-Resolution The introduction of SRCNN [8] pioneered end-to-end convolutional architectures for mapping LR images to their HR counterparts, substantially outperforming classical interpolation. FSRCNN [9] and ESPCN [31] improved efficiency by

moving the upsampling step into the network via deconvolution and sub-pixel convolutions, respectively. SRResNet [18] and EDSR [21] refined residual designs to have an exceptional performance in SR. EDSR removed BatchNorm [16] layers and stacked more residual blocks to boost fidelity. RCAN [35] further enhanced performance through channel-attention layers that dynamically reweight feature maps. After the advent of vision transformer, ViT-based models have been also applied to super-resolution tasks. SwinIR [20] adapts the Swin Transformer for SR by stacking Residual Swin Transformer Blocks (RSTBs), each using windowed self-attention and cyclically shifted windows to efficiently capture both local textures and cross-window context. This design reduces computational complexity from $O\big((HW)^2\big)$ to $O\big(HW\,M^2\big)$ while still matching or surpassing full-attention ViT restorers. However, large-scale models employ very deep architectures with millions of parameters, rendering them impractical for mobile or real-time applications.

Contrastive Learning. Contrastive learning trains representations by pulling semantically related views together while pushing apart other samples, typically via the InfoNCE loss [26]. Seminal works such as SimCLR [7] employ large batch sizes and random augmentations, whereas MoCo [12] uses a momentum encoder with a memory bank to efficiently harvest negatives. More recently, contrastive objectives have been adopted in knowledge distillation in super-resolution tasks. Distilling statistical information of the intermediate feature maps [13, 22]. CSD [33] align teacher–student embeddings via InfoNCE, demonstrating richer relational transfer and superior compactness compared to traditional L_1/L_2 feature matching. FACD [25] selectively applies feature contrastive learning by comparing output patches of the student and teacher network. These methods show that embedding-based relational cues can yield more robust and efficient student models.

KD for Super-Resolution. Early attempts to compress SISR models largely adopted vanilla KD techniques [14] from classification—matching pixel-wise outputs or intermediate feature maps [27]. These approaches typically yield only modest PSNR gains or even degrade visual quality, since the teacher's super-resolved output is itself an approximation of the ground truth. FAKD [13] align second-order feature affinities by injecting structural priors. CSD [32] uses a frozen VGG network to impose contrastive bounds in a fixed feature space. While these improve stability over raw L_1 matching, they rely on pre-trained backbones or handcrafted similarity metrics, and they lack explicit control over both global and local relational features. DUKD [36] sidesteps feature matching by upcycling in-domain LR–HR pairs, enforcing label consistency under data augmentations. MiPKD [19] introduces multi-granularity and adapter modules to fuse teacher priors at various network depths. These methods improve robustness or high-frequency recovery.

3. Methodology

3.1. Preliminaries and Notations

We denote by $I^{\mathrm{LR}} \in \mathbb{R}^{H\times W\times 3}$ a low-resolution (LR) input image and by $I^{\mathrm{GT}} \in \mathbb{R}^{sH\times sW\times 3}$ its corresponding high-resolution (HR) ground truth. A pre-trained and frozen teacher network $\mathcal{F}_T$ and a lightweight student network $\mathcal{F}_S$ produce output images:

$$I^T_{SR} = \mathcal{F}_T\big(I^{\mathrm{LR}}\big), \quad I^S_{SR} = \mathcal{F}_S\big(I^{\mathrm{LR}}\big) \tag{1}$$

We also employ a discriminator D for adversarial guidance, and two projection heads ϕ_{img} and ϕ_{patch} for contrastive embedding. I^T_{SR} and I^S_{SR} are output images of the teacher and student models, respectively.

3.2. Motivation

Classical knowledge distillation [14] uses the "logits" of the networks outputs by applying a softmax over them and matching the resulting probability distribution for tasks such as classification and detection. This is generally not effective and rather provide detrimental results [4]. Furthermore, in super resolution, the network's output is a full HR image or a dense tensor, not a probability vector over a fixed size. Logits for example, in a classification task collapse all spatial information into a single vector where SR tasks must recover. The reconstruction and response-based terms alone suffer from inheriting the teacher's artifacts: pixel-wise L1 matching cannot distinguish between correct details and teacher errors. Thus, the "dark knowledge" from the teacher model $\mathcal{F}_T$ outputs inaccurate approximation of the ground-truth distributions of HR image, I_{HR} [36].

To overcome these major issues, we propose two main components: multi-scale contrastive loss and adversarial distillation loss components. Multi-scale contrastive distillation remedies this by teaching the student how different regions relate, both at the whole-image and fine-patch levels, so that structural consistency and textural fidelity are explicitly enforced in a shared embedding space. This relational learning goes beyond second-order or frequency-band alignment or simple intermediate feature distillation, because it directly models both global semantics and local details. Meanwhile, the adversarial term injects a learned perceptual prior: by forcing the student to "fool" a discriminator trained on real HR images, we enable the student to generate textures and sharpness that can exceed the

teacher's own perceptual ceiling. Together, contrastive and adversarial guidance balance stability (from L1 and InfoNCE losses) with high-quality, realistic image generation. Details of each component will be elaborated in the latter sections.

3.3. Multi-scale Contrastive Learning

Previous work that leverages contrastive learning, CSD [32], tightly couples student and teacher via channel-splitting. It uses a frozen VGG-19 network to extract multi-layer feature maps and then applies an L1-based contrastive ratio between teacher vs. student and negatives in that fixed feature space. In comparison, our contrastive loss first learns two small projection heads (one for whole-image, one for patches) that map directly from SR outputs into a dedicated embedding space. These projection heads are trained end-to-end, so your distillation space adapts specifically to the SR task rather than relying on generic ImageNet [10] features. Furthermore, our method treats student and teacher as fully separate networks. We can plug the distillation losses into any off-the-shelf SR backbone without modifying its channel configuration, making it broadly applicable. Lastly, our contrastive learning method leverages multi-scale relational alignment enforces both image-level and patch-level InfoNCE objectives. This multi-scale contrast ensures that the student learns both the global structure and fine textures.

Image-level InfoNCE Loss. Let $z_T = \phi_{\text{img}}(I^T_{SR})$ and $z_S = \phi_{\text{img}}(I^S_{SR})$ where z_T and z_S are contrastive embeddings of the teacher's and student's super–resolved outputs. The projection head ϕ_{img} consists of a 3×3 convolution (mapping $3 \to d/2$ channels), ReLU, global average pooling, flattening, and a fully-connected layer ($d/2 \to d$), followed by L_2–normalization. These d–dimensional vectors lie in a learned latent space where cosine similarity reflects structural and perceptual alignment. The image-level InfoNCE loss is:

$$\mathcal{L}_{\text{img}} = -\log \frac{\exp\big(\text{sim}(z_S,\, z_T)/\tau\big)}{\sum_k \exp\big(\text{sim}(z_S,\, z_k)/\tau\big)} \tag{2}$$

where $\{z_k\}$ are negatives from the batch and τ is a temperature. τ is initialized as 0.07.

Patch-Level Contrastive Loss To enforce fine-grained textural consistency, we randomly extract $N = 8$ patches of size 48×48 from each super–resolved output, denoted $\{p^i_T\}$ for the teacher and $\{p^i_S\}$ for the student. Each patch is passed through the patch-level projection head ϕ_{patch}:

$$u^i_T = (\phi_{\text{patch}}(p^i_T)), \quad u^i_S = (\phi_{\text{patch}}(p^i_S)) \tag{3}$$

where u^i_T and u^i_S denotes teacher and student patch embeddings, respectively. We then apply an InfoNCE objective over these local embeddings:

$$\mathcal{L}_{\text{patch}} = -\sum_{i=1}^{N} \log \frac{\exp\big(\text{sim}(u^i_S,\, u^i_T)/\tau\big)}{\sum_{j=1}^{N} \exp\big(\text{sim}(u^i_S,\, u^j_T)/\tau\big)} \tag{4}$$

Where sim denotes the cosine similarity between two d-dimensional vectors. Ths function explicitly pulls each student patch embedding u^i_S toward its corresponding teacher patch u^i_T and pushes it away from all other patches $\{u^j_T\}_{j \neq i}$, thereby focusing the student on local structural and textural patterns. Compared to a standard global contrastive loss, the patch-level objective captures localized details such as edges and patterns that are often lost when averaging over the full image. Unlike pixel-wise L_1 or feature-affinity KD methods which treat all pixels or feature channels uniformly, patch-level contrast enforces discriminative local relationships, making the student more robust to spatially varying artifacts. Furthermore, our patch-level loss directly aligns local distributions in a learned embedding space, yielding sharper, more realistic textures without requiring specialized frequency decompositions or auxiliary data mining. Details are shown in Figure 2.

3.4. Adversarial Distillation

Incorporating an adversarial loss into our distillation framework provides a powerful perceptual prior that directly models the HR image manifold rather than merely mimicking the teacher's imperfect outputs. Concretely, we train a discriminator D with

$$\mathcal{L}_D \;=\; -\,\mathbb{E}_{I^{\text{GT}}}\big[\log D(I^{\text{GT}})\big] \;-\; \mathbb{E}_{\hat{I}_S}\big[\log\big(1-D(\hat{I}_S)\big)\big], \tag{5}$$

and update the student to minimize

$$\mathcal{L}_{\text{adv}} \;=\; -\,\mathbb{E}_{\hat{I}_S}\big[\log D(\hat{I}_S)\big]. \tag{6}$$

This adversarial term differs from traditional pixel- or feature-level KD losses by enabling the student to generate fine-grained textures and break through the teacher's perceptual ceiling. GAN objective requires no auxiliary data manipulation or handcrafted transforms—its learned discriminator automatically captures natural image statistics and focuses the student on high-frequency realism. While GAN training introduces additional parameters and potential instability, when properly balanced it yields sharper edges, and richer texture details.

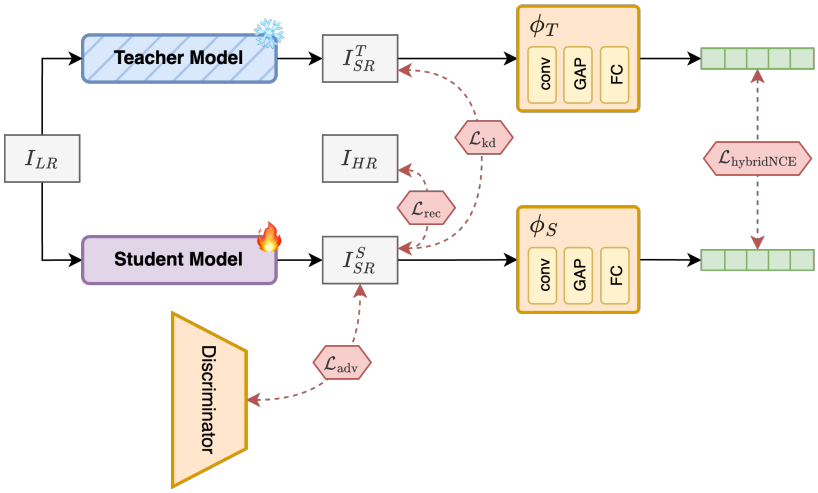

Figure 1. Overall framework of MCAD-KD. HybridNCE denotes the multi-scale contrastive loss.

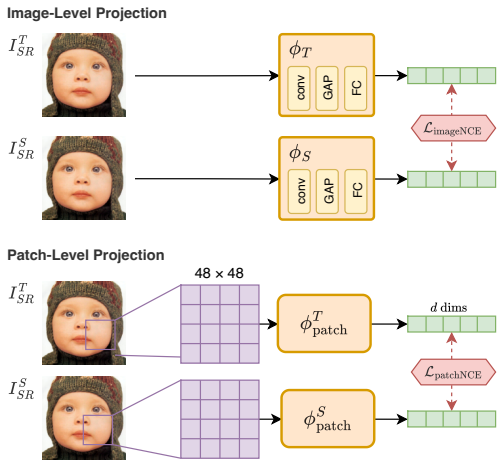

Figure 2. Detailed figures of image-level projection and patch-level projection.

Overall Loss Function. Our training objective is a weighted sum of four losses:

$$\mathcal{L}_{\text{total}} = \lambda_{rec}\mathcal{L}_{\text{rec}} + \lambda_{resp}\mathcal{L}_{\text{resp}} + \lambda_{cont}\mathcal{L}_{\text{cont}} + \lambda_{adv}\mathcal{L}_{\text{adv}}.$$

These are defined as:

$$\mathcal{L}_{\text{rec}} = \|I_{SR}^{S} - I^{\text{GT}}\|_1, \tag{7b}$$

$$\mathcal{L}_{\text{resp}} = \|I_{SR}^{S} - I_{SR}^{T}\|_1, \tag{7c}$$

$$\mathcal{L}_{\text{cont}} = \lambda_{\text{img}}\,\mathcal{L}_{\text{img}} + \lambda_{\text{patch}}\,\mathcal{L}_{\text{patch}}, \tag{7d}$$

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\hat{I}_S}\left[-\log D(\hat{I}_S)\right] \tag{7e}$$

Here, $\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{resp}}, \mathcal{L}_{\text{cont}}$ and $\mathcal{L}_{\text{adv}}$ denote the reconstruction, response-based distillation, multi-scale contrastive, and adversarial losses, respectively, as defined in Eqs. (7b)–(7e).

Table 1. Hyperparameter settings used throughout our experiments.

Term	Symbol	Value
Reconstruction weight	λ_{rec}	3.0
Response-KD weight	λ_{resp}	0.5
Contrastive total weight	λ_{cont}	0.05
Image-level weight	λ_{img}	0.3
Patch-level weight	λ_{patch}	0.7
Patch size	—	48×48
# of patches	N	8
Adversarial weight	λ_{adv}	0.01

4. Experimental Results

4.1. Experimental Settings

Backbones and Evaluation. We use EDSR [21], RCAN [35] and SwinIR [20] as backbone models to evaluate the performance of our MCAD-KD. We compare our method with recent works of knowledge distillation including Vanilla-KD [14], RKD [28], AT [], FitNet [29], FAKD [13], CSD [32], DUKD [36] and MiPKD [19]. We calculate the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) on the Y channel of the YCbCr color space. We use 800 images from DIV2K [1] dataset. We evaluated our work under four classical benchmark dataset, Set5 [3], Set14 [34], BSD100 [24] and Urban100 [15]. The SR network specification of teacher and student models is illustrated in Table 2.

Table 2. Specifications and computational statistics of SR models at ×4 scale. FLOPs and parameter counts are evaluated using 3×256×256 input images on an A100 GPU with 80GB VRAM. Block counts refer to residual blocks in EDSR and RCAN (per residual group) and Swin transformer blocks in SwinIR.

Model	Role	Network			FLOPs (G)	#Params
		Channel	Block	Group		
EDSR [21]	Teacher	256	32	-	3293.35	43.09 M
	Student	64	32	-	207.28	2.70 M
RCAN [35]	Teacher	64	20	10	1044.03	15.59 M
	Student	64	6	10	366.98	5.17 M
SwinIR [20]	Teacher	180	6	-	861.27	11.90 M
	Student	60	4	-	121.48	1.24 M

Training Details. All models are trained with Adam [17] optimizer with $\beta_1 = 0.9, \beta = 0.99$ and $\epsilon = 10^-8$ with a batch size of 32 and a total of 2.5×10^5 iterations. The initial learning rate was set to 10^{-4} for EDSR, RCAN and SwinIR. The learning rate is decayed by a factor of 10 at every 10^5 iterations. We set the hyperparameter of λ_{rec}, λ_{respkd}, λ_{cont}, λ_{adv} as $3.0, 0.5, 0.05$ and 0.01, respectively. The total contrastive loss weight is divided into image-level and patch-level weight. The values λ_{img} and λ_{patch} are 0.3 and 0.7, respectively. Overall summary of the hyperparameters can be found in Table 1.

5. Results Analysis

Comparison with Baseline Methods. Quantitative results for EDSR [21], RCAN [35], SwinIR [20] of scales $\times 2, \times 3, \times 4$ is presented in Table 3, 4, 5. From the results, we observe that MCAD-KD consistently delivers the largest PSNR/SSIM gains over both the scratch student and every competing distillation method, but the magnitude of those gains varies with architectures. On EDSR, we observe improvements of +0.67 dB at ×4(31.96 to 32.63 dB) and +0.64 dB at ×3 (27.99 to 28.63 dB) on Urban100, whereas RCAN sees a more modest +0.24 dB at ×2 and +0.20 dB at ×3. SwinIR, despite already benefiting from efficient windowed attention, still gains roughly +0.60 dB on Urban100 at ×2, underscoring MCAD-KD's ability to refine Transformer outputs. We believe that these differences are directly from mostly the multi-scale contrastive loss. Multi-scale contrastive distillation transfers both global structure and fine-patch relations; deep models like EDSR and SwinIR, with their larger capacity and richer feature hierarchies, can more fully exploit these relational cues, producing larger PSNR gains. In comparison, RCAN's built-in channel-attention already captures some global–local dependencies, so the marginal benefit of extra contrastive supervision is smaller. Adversarial distillation guidance also helps in marginal benefits, injecting a learned perceptual prior that sharpens textures. Our patch-level InfoNCE loss directly supervises that process by sampling 48×48 patches and forcing the student to match the teacher's patch embeddings in a learned space. In effect, we give SwinIR an explicit "local-texture teacher" signal on top of its internal attention, which sharpens repetitive structures. Especially, in Urban100 where skyscrapers, facades, and tiled streets exhibit strong, repeating high-frequency patterns, the perceptual prior of using the adversarial loss pushes aligned features toward the true HR distribution's high-frequency statistics. This produces marginal but visually significant improvements : crisper edges, more realistic textures, and a further PSNR gain of 0.05–0.10 dB over contrastive-only distillation.

5.1. Qualitative Analysis

Figure 3 illustrates representative ×4 super-resolution crops from Urban100, comparing the student trained from scratch, with DUKD, MiPKD, CSD, and our MCAD-KD. Scratch and vanilla KD outputs exhibit pronounced blurring on repetitive patterns (e.g. window frames and brick textures), while DUKD and MiPKD slightly improve sharpness but still suffer from washed-out edges and inconsistent local details. CSD recovers more structure via its VGG-based contrastive bound, yet fine textures remain oversmoothed. In contrast, MCAD-KD produces markedly crisper edges, faithfully reconstructs small-scale patterns such as the grid of window panes and the weave of textured walls, and suppresses halo artifacts around high-contrast boundaries. Our patch-level InfoNCE term explicitly guides the student to match the teacher's local textural statistics, and the adversarial loss injects a learned perceptual prior that recovers plausible high-frequency detail. Together, these components yield SR outputs that not only score highest in PSNR/SSIM but also look visually closest to the ground truth, especially on the highly structured

Table 3. Quantitative comparison (average PSNR/SSIM) between MCAD-KD and other distillation methods for EDSR of three SR scales. The best and second-best performances are highlighted in bold and underlined, respectively. The EDSR teacher model c256b32 is distilled to c64b32 student model.

Scale	Method	Set5	Set14	BSD100	Urban100
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
×2	Scratch	38.00/0.9605	33.57/0.9171	32.17/0.8996	31.96/0.9268
	Vanilla-KD	38.04/0.9606	33.58/0.9172	32.19/0.8998	31.98/0.9269
	FitNet	37.59/0.9589	33.09/0.9136	31.79/0.8953	30.46/0.9111
	RKD	38.03/0.9606	33.57/0.9173	32.18/0.8998	31.96/0.9270
	FAKD	37.99/0.9606	33.60/0.9173	32.19/0.8998	32.04/0.9275
	CSD	38.06/0.9607	33.65/0.9179	32.22/0.9004	32.26/0.9300
	DUKD	38.15/0.9610	33.80/0.9195	32.27/0.9007	32.53/0.9320
	MiPKD	38.16/**0.9611**	33.85/0.9194	32.27/0.9008	32.52/0.9318
	MCAD-KD	**38.19/0.9611**	**33.88/0.9203**	**32.31/0.9013**	**32.63/0.9329**
×3	Scratch	34.39/0.9270	30.32/0.8417	29.08/0.8046	27.99/0.8489
	Vanilla-KD	34.43/0.9273	30.34/0.8422	29.10/0.8050	28.00/0.8491
	FitNet	33.35/0.9178	29.71/0.8323	28.62/0.7949	26.61/0.8167
	RKD	34.43/0.9274	30.33/0.8423	29.09/0.8051	27.96/0.8493
	FAKD	34.39/0.9272	30.34/0.8426	29.10/0.8052	28.07/0.8511
	CSD	34.45/0.9275	30.32/0.8430	29.11/0.8061	28.21/0.8549
	DUKD	34.59/0.9287	30.47/0.8448	29.20/0.8073	28.44/0.8578
	MiPKD	34.59/0.9287	30.48/0.8447	29.19/0.8070	28.41/0.8571
	MCAD-KD	**34.65/0.9290**	**30.54/0.8457**	**29.24/0.8082**	**28.63/0.8633**
×4	Scratch	32.29/0.8965	28.68/0.7840	27.64/0.7380	26.21/0.7893
	Vanilla-KD	32.30/0.8965	28.70/0.7842	27.64/0.7382	26.21/0.7897
	FitNet	31.65/0.8873	28.33/0.7768	27.38/0.7309	25.40/0.7637
	RKD	32.30/0.8965	28.69/0.7842	27.64/0.7383	26.20/0.7899
	FAKD	32.27/0.8960	28.65/0.7836	27.62/0.7379	26.18/0.7895
	CSD	32.34/0.8974	28.72/0.7856	27.68/0.7396	26.34/0.7948
	DUKD	32.47/0.8981	28.80/0.7866	27.71/0.7403	26.45/0.7963
	MiPKD	32.46/0.8981	28.79/0.7863	27.71/0.7400	26.45/0.7960
	MCAD-KD	**32.50/0.8984**	**28.84/0.7870**	**27.76/0.7407**	**26.51/0.7967**

scenes in Urban100.

5.2. Ablation Studies

Ablation on Contrastive vs. Adversarial Distillation. Table 6 demonstrates the ablation analysis of absence of each loss. When distilling EDSR at ×4 scale on Urban100, using only the multi-scale contrastive loss yields a PSNR/SSIM of 26.48/0.7963, whereas using only the adversarial loss gives 26.35/0.7948. Combining both losses boosts performance to 26.51/0.7967. This demonstrates that contrastive supervision is the primary driver of reconstruction accuracy by aligning relational cues in a learned embedding space while the adversarial term provides a complementary perceptual refinement, sharpening textures and pushing the student beyond the teacher's inherent ceiling.

Ablation on Image vs. Patch-Level Contrastive Loss. Table 6 shows the ablation study of using a single-scale contrastive loss. Isolating the two scales of our contrastive distillation on EDSR reveals that the image-level term alone achieves 26.49/0.7966 and the patch-level term alone 26.48/0.7966. Leveraging multi-scale each capture distinct aspects of the SR task: the former preserves overall semantics, while the latter enforces fine-grained detail.

Ablation on Multi-Scale Contrastive for SwinIR. Table 8 shows the ablation study of using a single-scale contrastive loss and the imapact of patch-level loss for SwinIR. For SwinIR at ×4 scale on Urban100, the patch-level contrastive loss alone (26.55/0.8002) outperforms the image-level variant (26.50/0.7997), reflecting SwinIR's architectural strength in capturing global context via windowed attention. Nonetheless, combining both scales still provides the best result (26.57/0.8006), indicating that even Transformer-based SR benefits from explicit local-texture supervision. The patch-level term

Table 4. Quantitative comparison (average PSNR/SSIM) between MCAD-KD and other distillation methods for RCAN of three SR scales. The best and second-best performances are highlighted in bold and underlined, respectively.

Scale	Method	Set5	Set14	BSD100	Urban100
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
×2	Scratch	38.13/0.9610	33.78/0.9194	32.26/0.9007	32.63/0.9327
	KD	38.18/0.9611	33.83/0.9197	32.29/0.9010	32.67/0.9329
	FitNet	37.97/0.9602	33.57/0.9174	32.19/0.8999	32.06/0.9279
	RKD	38.18/0.9612	33.78/0.9191	32.29/0.9011	32.70/0.9330
	FAKD	38.17/0.9612	33.83/0.9199	32.29/0.9011	32.65/0.9330
	DUKD	38.23/0.9614	33.90/0.9201	32.33/0.9016	32.87/0.9349
	MiPKD	38.21/0.9613	33.92/0.9203	32.32/0.9015	32.83/0.9344
	MCAD-KD	38.24/0.9613	33.90/0.9202	32.35/0.9016	32.90/0.9348
×3	Scratch	34.61/0.9288	30.45/0.8444	29.18/0.8074	28.59/0.8610
	KD	34.61/0.9291	30.47/0.8447	29.21/0.8080	28.62/0.8612
	FitNet	34.21/0.9248	30.20/0.8399	29.05/0.8044	27.89/0.8472
	RKD	34.67/0.9292	30.48/0.8451	29.21/0.8080	28.60/0.8610
	FAKD	34.63/0.9290	30.51/0.8453	29.21/0.8079	28.62/0.8612
	DUKD	34.74/0.9296	30.54/0.8458	29.25/0.8088	28.79/0.8646
	MiPKD	34.72/0.9296	30.55/0.8458	29.25/0.8087	28.76/0.8640
	MCAD-KD	34.75/0.9296	30.55/0.8458	29.29/0.8090	28.81/0.8650
×4	Scratch	32.31/0.8966	28.69/0.7842	27.64/0.7384	26.37/0.7949
	KD	32.45/0.8980	28.76/0.7860	27.67/0.7400	26.49/0.7980
	FitNet	31.99/0.8899	28.50/0.7789	27.55/0.7353	25.90/0.7791
	RKD	32.39/0.8974	28.74/0.7856	27.67/0.7399	26.47/0.7981
	FAKD	32.46/0.8980	28.77/0.7860	27.68/0.7400	26.50/0.7980
	DUKD	32.56/0.8990	28.83/0.7870	27.72/0.7410	26.62/0.8020
	MiPKD	32.46/0.8982	28.77/0.7860	27.69/0.7402	26.55/0.7998
	MCAD-KD	32.57/0.8994	28.84/0.7875	27.74/0.7413	26.71/0.8031

Table 5. Quantitative comparison (average PSNR/SSIM) between MCAD-KD and other applicable distillation methods for SwinIR of three SR scales. The best and second-best performances are highlighted in bold and underlined, respectively.

Scale	Method	Set5	Set14	BSD100	Urban100
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
×2	Scratch	38.01/0.9607	33.57/0.9178	32.19/0.9000	32.05/0.9279
	KD	38.04/0.9608	33.61/0.9184	32.22/0.9003	32.09/0.9282
	DUKD	38.13/0.9610	33.78/0.9194	32.26/0.9007	32.63/0.9327
	MiPKD	38.14/0.9611	33.76/0.9194	32.29/0.9011	32.46/0.9313
	MCAD-KD	38.18/0.9614	33.80/0.9196	32.32/0.9015	32.7/0.9333
×3	Scratch	34.41/0.9273	30.43/0.8437	29.12/0.8062	28.20/0.8537
	KD	34.44/0.9275	30.45/0.8443	29.14/0.8066	28.23/0.8545
	DUKD	34.55/0.9285	30.53/0.8456	29.20/0.8080	28.53/0.8604
	MiPKD	34.53/0.9283	30.52/0.8456	29.19/0.8079	28.47/0.8591
	MCAD-KD	34.58/0.9289	30.57/0.8470	29.22/0.8088	28.68/0.8635
×4	Scratch	32.31/0.8955	28.67/0.7833	27.61/0.7379	26.15/0.7884
	KD	32.27/0.8954	28.67/0.7833	27.62/0.7380	26.15/0.7887
	DUKD	32.41/0.8973	28.79/0.7860	27.69/0.7405	26.43/0.7972
	MiPKD	32.39/0.8971	28.76/0.7854	27.68/0.7403	26.37/0.7956
	MCAD-KD	32.47/0.8988	28.84/0.7880	27.72/0.7419	26.57/0.8006

corrects SwinIR's tendency to oversmooth, while the image-level term ensures consistency in large-scale structures.

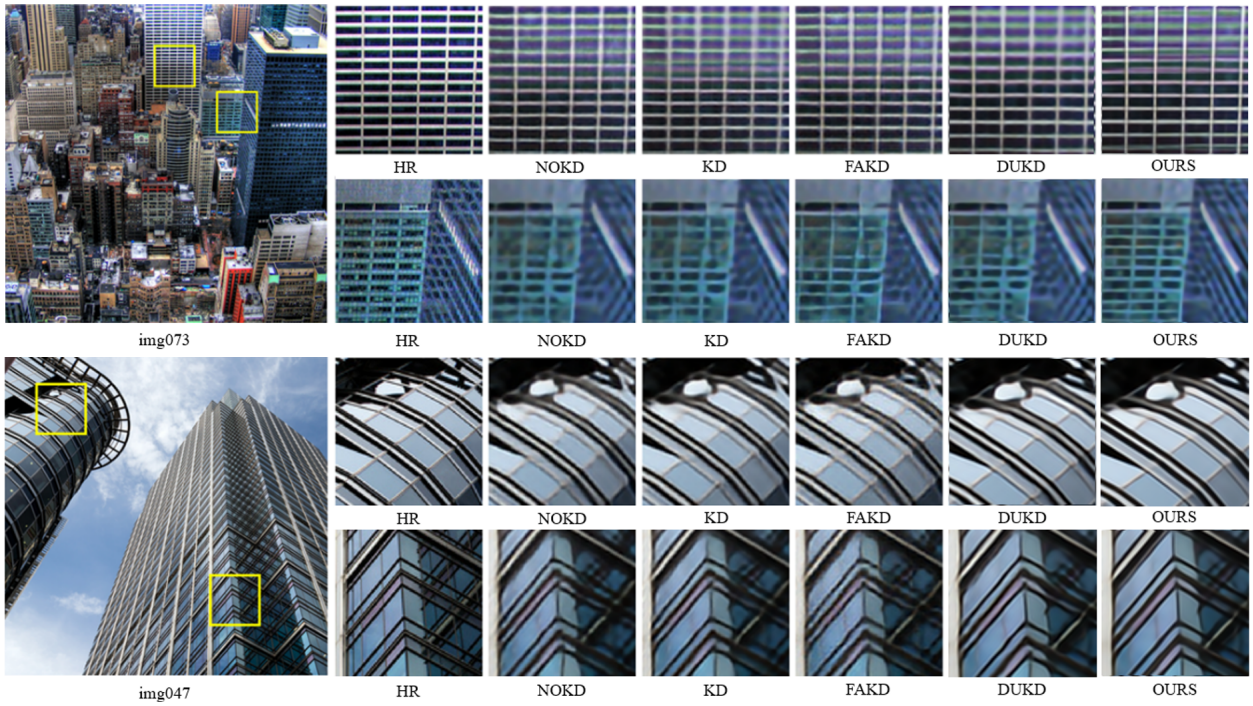

Figure 3. Qualitative outputs of ×4 SR examples of EDSR models with different KD methods. Img073 and img047 is taken from Urban100.

Table 6. Ablation on the two modules of MCAD-KD of $\times 4$ scale. AD Loss stands for Adversarial Distillation loss.

Contrastive Loss	AD Loss	Urban100 PSNR / SSIM
✓	×	26.48 / 0.7963
×	✓	26.35 / 0.7948
✓	✓	26.51 / 0.7967

Table 7. Ablation on the two parts of multi-scale contrastive loss for$\times 4$ scale of EDSR . The last row indicate multi-scale contrastive loss.

Image-level	Patch-level	Urban100 PSNR / SSIM
✓	×	26.49 / 0.7966
×	✓	26.48 / 0.7966
✓	✓	26.51 / 0.7967

Table 8. Ablation on the two parts of multi-scale contrastive loss for$\times 4$ scale of SwinIR .

Image-level	Patch-level	Urban100 PSNR / SSIM
✓	×	26.50 / 0.7997
×	✓	26.55 / 0.8002
✓	✓	26.57 / 0.8006

6. Conclusion

We have presented MCAD-KD, a unified distillation framework for single-image super-resolution that synergistically combines multi-scale contrastive learning with adversarial guidance. By aligning global and local embeddings via learned projection heads and injecting a perceptual prior through a adversarial loss, MCAD-KD consistently outperforms prior KD methods across EDSR, RCAN, and SwinIR backbones particularly on the structurally rich Urban100 dataset. Our plug-and-play design requires no architectural modifications and scales, paving the way for efficient, high-fidelity SR on resource-constrained devices.

Acknowledgements. This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the Global Scholars Invitation Program (RS-2024-00459638) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), the Basic Science Research Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. IIPT-2025-RS-2024-00346737) and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. IITP-2025-RS-2023-00254129, Graduate School of Metaverse Convergence (Sungkyunkwan University)).

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 4

[2] Sajid Anwar, Kyuyeon Hwang, and Wonyong Sung. Structured pruning of deep convolutional neural networks. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 13(3):1–18, 2017. 1

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 4

[4] Xinning Chai, Yao Zhang, Yuxuan Zhang, Zhengxue Cheng, Yingsheng Qin, Yucai Yang, and Li Song. Distillation-supervised convolutional low-rank adaptation for efficient image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1431–1440, 2025. 2

[5] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 1

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 2

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1

[9] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016. 1

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 3

[11] Robert M. Gray and David L. Neuhoff. Quantization. *IEEE transactions on information theory*, 44(6):2325–2383, 1998. 1

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[13] Zibin He, Tao Dai, Jian Lu, Yong Jiang, and Shu-Tao Xia. Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution. In *2020 IEEE international conference on image processing (ICIP)*, pages 518–522. IEEE, 2020. 2, 4

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 4

[15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 4

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 2

[17] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, et al. Photo-realistic single image super-resolution using a generative adversarial network [c]. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[19] Simiao Li, Yun Zhang, Wei Li, Hanting Chen, Wenjia Wang, Bingyi Jing, Shaohui Lin, and Jie Hu. Knowledge distillation with multi-granularity mixture of priors for image super-resolution. *arXiv preprint arXiv:2404.02573*, 2024. 2, 4

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1, 2, 4, 5

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1, 2, 4, 5

[22] Cencen Liu, Dongyang Zhang, and Ke Qin. Knowledge distillation for single image super-resolution via contrastive learning. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1079–1083, 2024. 2

[23] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2604–2613, 2019. 1

[24] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, pages 416–423. IEEE, 2001. 4

[25] Hyeon-Cheol Moon, Jae-Gon Kim, Jinwoo Jeong, and Sungjei Kim. Feature-domain adaptive contrastive distil-

lation for efficient single image super-resolution. *IEEE Access*, 11:131885–131896, 2023. 2

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[27] SeongUk Park and Nojun Kwak. Local-selective feature distillation for single image super-resolution. *arXiv preprint arXiv:2111.10988*, 2021. 2

[28] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019. 4

[29] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 1, 4

[30] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 1

[31] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 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*, pages 1874–1883, 2016. 1

[32] Yanbo Wang, Shaohui Lin, Yanyun Qu, Haiyan Wu, Zhizhong Zhang, Yuan Xie, and Angela Yao. Towards compact single image super-resolution via contrastive self-distillation. *arXiv preprint arXiv:2105.11683*, 2021. 2, 3, 4

[33] Yanbo Wang, Shaohui Lin, Yanyun Qu, Haiyan Wu, Zhizhong Zhang, Yuan Xie, and Angela Yao. Towards compact single image super-resolution via contrastive self-distillation. *arXiv preprint arXiv:2105.11683*, 2021. 2

[34] Roman Zeyde, Michael Elad, and Matan Protter. 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*, pages 711–730. Springer, 2012. 4

[35] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 1, 2, 4, 5

[36] Yun Zhang, Wei Li, Simiao Li, Hanting Chen, Zhijun Tu, Wenjia Wang, Bingyi Jing, Shaohui Lin, and Jie Hu. Data upcycling knowledge distillation for image super-resolution. *arXiv preprint arXiv:2309.14162*, 2023. 1, 2, 4