AMID: ATTRIBUTION MATCHING AND INVERSE-NETWORK DISTILLATION FOR HETEROGENEOUS SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distilling image super-resolution (SR) models across heterogeneous architectures is challenging because their hidden representations are not directly comparable, while output matching alone does not transfer how the teacher exploits low-resolution evidence. We propose **Attribution Matching and Inverse-Network Distillation (AMID)**, which uses teacher outputs and output gradients **without accessing teacher intermediate features**. Global Attribution Map Distillation (GAMD) matches integrated-gradient attribution maps of guidance-weighted structural responses, **transferring the teacher's low-resolution pixel utilization and stabilizing reconstruction fidelity**. Inverse Self-Consistency Distillation (ISCD) passes teacher and student outputs through a frozen inverse network architecturally isomorphic to the student, enabling multi-layer contrastive alignment in a shared, degradation-aware proxy space. An inverse reconstruction constraint further enforces degradation consistency. Experiments with five heterogeneous teachers across five benchmarks demonstrate improvements over representative distillation methods at $2\times$, $3\times$, and $4\times$. At $3\times$ with a SwinIR teacher, AMID outperforms all compared baselines across all eight metrics and improves MUSIQ by 1.22 over the strongest competing method. All auxiliary components are used **only during training**. AMID preserves the student's architecture and **inference cost**, making it practical for deployment-oriented heterogeneous SR distillation.

1 INTRODUCTION

Image super-resolution (SR) reconstructs high-resolution images from low-resolution observations. CNN-based methods achieve strong reconstruction performance through residual and dense feature learning (Zhang et al., 2018b;c; Wang et al.; Zhang et al., 2022). Transformer models improve long-range dependency modeling (Vaswani et al., 2017; Liang et al., 2021; Chen et al., 2025), while state-space architectures provide global receptive fields with favorable complexity (Liu et al., 2024; Guo et al., 2024b; 2025). Diffusion-based approaches further advance perceptual quality (Wang et al., 2024; Yu et al., 2024; Lin et al., 2025; Li et al., 2025c; Yi et al., 2025; Li et al., 2025b; Fei et al., 2025). Despite these advances, hardware support often determines which of these architectures can be deployed, motivating distillation from **a stronger teacher into a student** the device can run.

In practice, teacher and student models are often heterogeneous, differing in architecture, feature dimensionality, and representation hierarchy. Their hidden states are therefore difficult to align directly. Output matching avoids this feature mismatch, but it supervises only the final high-resolution (HR) reconstruction and does not transfer how the teacher exploits low-resolution (LR) evidence. This raises a practical question: can a student inherit reconstruction behavior from a heterogeneous teacher **without reading or aligning teacher hidden features**? We study this **hidden-feature-free** setting, where teacher intermediate activations are not accessed, while teacher predictions and input-output gradients for attribution are available during training. Our goal is to transfer reconstruction behavior without architecture-specific teacher features or modifications to the deployed student.

Existing SR distillation methods commonly require homogeneous architectures or alignable intermediate features (Xie et al., 2023; Guo et al., 2024a; Xiao et al., 2022; Yoon et al., 2023). Heterogeneous methods use relational modeling or feature projectors to reduce representation mismatch (Park et al., 2019; Zhao et al., 2022; Zhang et al., 2024a), but still depend on hidden representations be-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

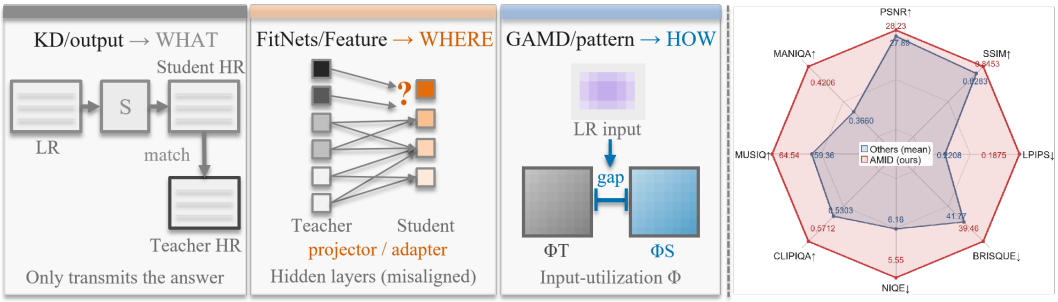


Figure 1: Left, an illustrative schematic of output KD (**what**), FitNets (**where**), and GAMD (**how**). Right, AMID versus the mean of other methods at 3× with a SwinIR teacher in Table 2.

tween incompatible spaces. Output-level methods avoid hidden-state matching (Wang et al., 2021b; Zhang et al., 2024b), yet do not explicitly transfer how the teacher utilizes the input. Architecture-specific bridges (Wang et al., 2025) are also difficult to generalize across model families.

To address these limitations, we introduce **AMID** (Attribution Matching and Inverse-Network Distillation), a hidden-feature-free distillation formulation with two complementary components for heterogeneous SR. GAMD (Global Attribution Map Distillation) matches integrated-gradient attribution maps of guidance-weighted structural responses, **transferring teacher LR-pixel utilization without hidden features**. A tail-selective guidance map emphasizes locations where attribution supervision is most informative. ISCD (Inverse Self-Consistency Distillation) passes teacher and student outputs through the same frozen inverse network, architecturally isomorphic to the student. Multi-layer contrastive alignment in this shared, degradation-aware proxy space **avoids matching heterogeneous hidden states**, while an inverse reconstruction loss enforces degradation consistency. GAMD mainly **stabilizes fidelity**, whereas ISCD **drives perceptual improvements**.

Fig. 1 contrasts three distillation targets. Output-level method transfers **what** the teacher predicts, as in KD (Hinton et al., 2015). FitNets (Romero et al., 2014) specifies **where** supervision is applied by matching projected hidden features. GAMD targets **how** the teacher uses LR evidence by matching guidance-weighted attribution maps. The right panel compares AMID with the other methods. We evaluate AMID with a fixed RRDB student and five teachers spanning Transformer, state-space, hybrid, and diffusion architectures. Across five benchmarks and three upscaling factors, AMID improves over representative distillation baselines. Our contributions are summarized as follows:

- We introduce GAMD, which **transfers teacher input-utilization patterns** through guidance-weighted attribution matching without requiring carefully designed hidden features or architecture-specific adapters for alignment.
- We propose ISCD, which performs multi-layer contrastive alignment in a frozen, student-isomorphic inverse network, providing degradation-aware supervision in a shared proxy space and **overcoming the representation mismatch** across heterogeneous architectures.
- AMID combines GAMD and ISCD: GAMD stabilizes fidelity, while ISCD improves perceptual quality. With five heterogeneous teachers on five benchmarks, it outperforms existing methods on **most metrics** at 2×/3×/4× scales, **without extra inference cost**.

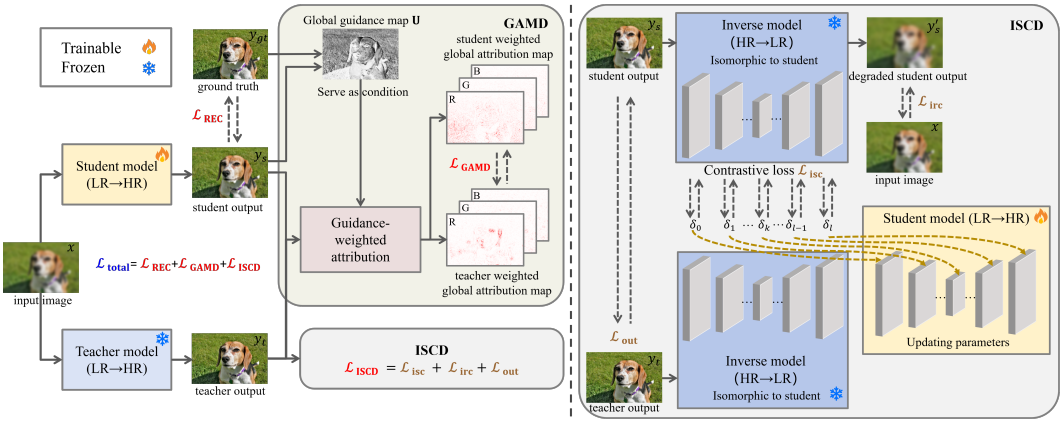
2 Related Work

Response-based distillation Knowledge distillation (KD) transfers softened teacher predictions to a student (Hinton et al., 2015). Subsequent methods improve response supervision through label regularization (Müller et al., 2019), teacher assistants (Mirzadeh et al., 2019), and intra- or inter-sample relations (Zhu et al., 2025). In SR, response-based approaches align reconstructed outputs using pixel-wise or perceptual objectives (Wang et al., 2021b; Zhang et al., 2024b; Son et al., 2026). These methods naturally support heterogeneous architectures because they avoid direct layer correspondence. However, output agreement does not explicitly transfer how the teacher uses LR evidence to form its prediction. Table 1 compares current SR distillation methods.

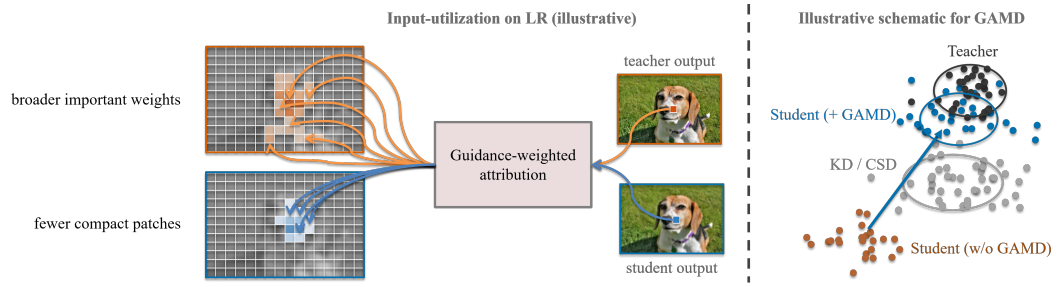
Intermediate feature-based distillation Romero et al. (2014) aligns hidden layers through projections, whereas Zagoruyko & Komodakis (2016) transfer matches spatial activations. Other methods transfer relations via Gram matrices (Yim et al., 2017), relational structures (Park et al., 2019),

Table 1: Comparison with representative distillation methods. Cross-arch: support for heterogeneous teacher-student architectures. Attribution: explicit attribution matching. Inverse proxy: alignment through a student-isomorphic inverse network. Training cost is reported qualitatively.

Method	Task	Cross-arch	Attribution	Inverse proxy	Training cost
KD (Hinton et al., 2015)	General	Yes	No	No	Very low
CSD (Wang et al., 2021b)	SR	Yes	No	No	Low
FAKD (He et al., 2020)	SR	No	No	No	Moderate
MiPKD (Li et al., 2024)	SR	No	No	No	High
MTKD (Jiang et al., 2024)	SR	Yes	No	No	High
DCKD (Zhou et al., 2024a)	SR	Yes	No	No	High
AugKD (Zhang et al., 2025)	SR	Yes	No	No	High
AMID (ours)	SR	**Yes**	**Yes**	**Yes**	Moderate



(a) **AMID** overview. GAMD matches guidance-weighted input attribution maps, while ISCD aligns teacher and student reconstructions through a frozen inverse network using contrastive, cycle, and output losses.



(b) Illustrative GAMD motivation (schematic). **Left:** guidance-weighted attribution analysis on LR input utilization (Teacher vs. Student patch weights). **Right:** conceptual view of the GAMD mechanism.

Figure 2: Overview of the proposed framework.

or contrastive objectives (Tian et al., 2019). He et al. (2020) distills channel-wise affinities for texture reconstruction. Such methods require compatible feature semantics or learned projectors (Hao et al., 2023; Li et al., 2025a; Wu et al., 2024), limiting transfer between distant architectures.

3 METHOD

As illustrated in Fig. 2a, a frozen teacher and trainable student map LR input x to HR reconstructions. GAMD weights each model's structural response with a guidance map derived from the student residual, then matches their integrated-gradient attribution maps. ISCD feeds both HR reconstructions into a shared frozen, student-isomorphic inverse network $\mathcal{I}$ and performs multi-layer contrastive alignment in the shared inverse-network feature space. Inverse reconstruction and output losses enforce cycle and prediction consistency. All auxiliary components are removed at inference.

3.1 GLOBAL ATTRIBUTION MAP DISTILLATION(**GAMD**)

Motivation. Conventional distillation (He et al., 2020; Wang et al., 2021b; Fang et al., 2023; Jiang et al., 2024; Hinton et al., 2015; Zhao et al., 2022; Passban et al., 2020; Guo et al., 2021; Zhang et al.,

2021) transfers outputs or features, conveying **what the teacher thinks the answer is** or **what the teacher sees in hidden layers**, but not **how the teacher uses input evidence**. Recent attribution methods (Lu et al., 2021; Wang et al., 2021a; Qiu et al., 2021; Xue et al., 2023; Lv et al., 2022; Yang et al., 2021; Crupi et al., 2024; Zhang et al., 2020; Anciukevicius et al., 2022; Lu et al., 2025) quantify input contributions to predictions. Local Attribution Map (LAM) (Sundararajan et al., 2017; Gu & Dong, 2020) uses Path-Integrated Gradients (PIG) to identify input regions contributing to an SR output. Building on this idea, GAMD constructs guidance-weighted structural responses and matches the resulting student attribution map to the teacher target (Fig. 2b), without accessing teacher hidden features or introducing cross-architecture adapters.

Preliminaries and formulation. The teacher's use of LR evidence is reflected in how its reconstruction depends on individual input pixels, which we characterize using attribution maps. Let $f \in \{\mathcal{T}, \mathcal{S}\}$ denote either the teacher or student SR model, with $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{sH\times sW\times C}$ for scale factor s. Given an LR input $x \in \mathbb{R}^{H\times W\times C}$, the corresponding ground truth is $y_{gt} \in \mathbb{R}^{sH\times sW\times C}$, where $C = 3$ for RGB images. Integrated Gradients (Sundararajan et al., 2017) requires a reference baseline x'. Following LAM (Gu & Dong, 2020), we obtain x' by Gaussian-blurring x, thereby reducing its high-frequency information while preserving the input dimensions.

To measure how input contributions accumulate from the baseline x' to the observed input x, let $\gamma(t)$ denote a Gaussian-blur path whose blur scale decreases continuously from the reference blur level at $t = 0$ to zero at $t = 1$. Thus, $\gamma(0) = x'$ and $\gamma(1) = x$. The path attribution from input element i to output element p is:

$$\mathrm{Att}_{p,i} = \int_0^1 \frac{\partial f_p(\gamma(t))}{\partial \gamma_i(t)} \frac{\partial \gamma_i(t)}{\partial t}\, dt. \tag{1}$$

This quantity integrates the sensitivity of output p to input i along the complete path. Further interpretation is provided in A.1.1 and all these definitions follow LAM (Gu & Dong, 2020). Treating all output locations equally can dilute the contribution of structurally distinctive regions. We therefore introduce a guidance map $\mathbf{U} \in [0, 1]$ derived from the student residual relative to the ground truth. Batch-wise min–max normalization produces $\hat{R} \in [0, 1]$, which represents relative residual magnitude within the current batch. Our inverted-Gaussian mapping assigns high weights to both residual extremes and low weights to the mid-range bulk. Thus, $\mathbf{U}$ serves as a deterministic saliency weight for attribution matching and is not an estimate of predictive uncertainty.

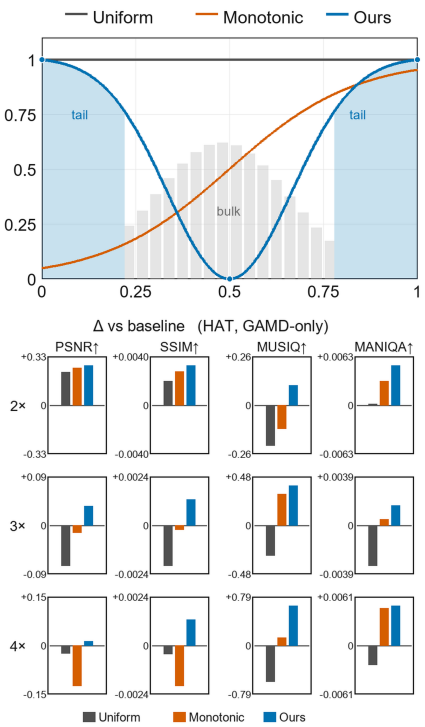


Figure 3: Uniform, Monotonic, and Ours maps $u(\hat{R})$ (top), with shaded tails and mid-range bulk. Signed Δ versus the $\mathcal{L}_{\mathrm{REC}}$-only baseline under HAT GAMD-only training (bottom).

Attribution guidance map estimation. The guidance map determines how strongly each HR location contributes to attribution matching. Given the student prediction $\mathbf{P} = \mathcal{S}(x) \in \mathbb{R}^{N\times C\times sH\times sW}$ and ground truth $\mathbf{G} = y_{gt}$, we compute the channel-aggregated residual as $R_{n,1,sh,sw} = \|\mathbf{P}_{n,:,sh,sw} - \mathbf{G}_{n,:,sh,sw}\|_2 = \sqrt{\sum_{c=1}^{C}(P_{n,c,sh,sw} - G_{n,c,sh,sw})^2}$. This produces $\mathbf{R} \in \mathbb{R}^{N\times 1\times sH\times sW}$, and the ℓ_2 rationale is in A.1.2.

We apply min-max normalization jointly over the batch and spatial dimensions: $\hat{R}_{n,1,sh,sw} = (R_{n,1,sh,sw} - R_{\min})/\max(R_{\max} - R_{\min}, \epsilon)$, where $R_{\min} = \min_{n,sh,sw} R_{n,1,sh,sw}$, $R_{\max} = \max_{n,sh,sw} R_{n,1,sh,sw}$, and $\epsilon > 0$ ensures numerical stability. This removes the absolute residual scale while preserving each value's relative position within the batch range. The guidance map $\mathbf{U} = u(\hat{R}) \in [0, 1]^{N\times 1\times sH\times sW}$ serves as a fixed weighting map during attribution computation. Fig. 3 compares uniform, monotonic, and inverted-Gaussian policies. Ours ($\mu = 0.5$, $\sigma = 1/6$) assigns low weight to the mid-range bulk and high

weight to both shaded tails. Its formulation is detailed in A.1.3, with the tail-selective rationale and controlled guidance-policy ablation provided in A.10 and Table 14. A spatial example is shown in Fig. 5. During GAMD, the map is held fixed and applied identically to both models, preventing gradients from propagating through the residual-based weighting.

GAMD loss. Directly matching $\mathrm{Att}_{p,i}$ for every HR output element is computationally expensive. We instead construct a scalar structural response for each model and attribute it to the LR input. For $m \in \{T, S\}$, the per-channel spatial-gradient magnitude is:

$$\mathcal{G}^m_{c,sh,sw}(x) = \sqrt{\Big(f^m_{c,sh,sw}(x) - f^m_{c,sh+1,sw}(x)\Big)^2 + \Big(f^m_{c,sh,sw}(x) - f^m_{c,sh,sw+1}(x)\Big)^2}, \quad (2)$$

where $1 \le sh < sH$ and $1 \le sw < sW$. We aggregate the four guidance values associated with each gradient location as: $\tilde{U}_{sh,sw} = U_{sh,sw} + U_{sh+1,sw} + U_{sh,sw+1} + U_{sh+1,sw+1}$. The resulting scalar response is:

$$\mathcal{R}^m(x; \mathbf{U}) = \sum_{c=1}^{C} \sum_{sh=1}^{sH-1} \sum_{sw=1}^{sW-1} \tilde{U}_{sh,sw}\, \mathcal{G}^m_{c,sh,sw}(x). \quad (3)$$

We apply equation 1 to the scalar response by replacing the output element f_p with $\mathcal{R}^m$. This removes the output index p and yields the input attribution.

$$\Phi^m_i(x) = \int_0^1 \frac{\partial \mathcal{R}^m(\gamma(t); \mathbf{U})}{\partial \gamma_i(t)} \frac{\partial \gamma_i(t)}{\partial t}\, dt, \qquad m \in \{T, S\}. \quad (4)$$

where $\mathbf{U}$ is held fixed throughout the integration path. Before matching, Φ^T and Φ^S are normalized channel-wise, and details on this normalization, path discretization, and optimization are provided in A.1.4. GAMD then minimizes their element-wise discrepancy:

$$\mathcal{L}_{\mathrm{GAMD}} = \frac{1}{|\Omega_x|} \sum_{i \in \Omega_x} \left| \bar{\Phi}^S_i(x) - \bar{\Phi}^T_i(x) \right|, \quad (5)$$

where Ω_x contains all channel and spatial elements of the LR, and $\bar{\Phi}$ denotes the normalized attribution map. Attribution-correspondence controls and visualizations are reported in A.11 and A.12.

3.2 Inverse self-consistency distillation(**ISCD**)

Motivation. Direct feature distillation is difficult when teacher and student models differ in feature dimensions and representation hierarchies. Projector-based methods (Li et al., 2025a; Wu et al., 2024) map incompatible features into a common space, but their alignment remains sensitive to semantic mismatch between selected layers (Tian et al., 2025). Output-mediated methods (Wang et al., 2021b; Zhang et al., 2024b) avoid teacher-feature matching by comparing reconstructed images through a VGG extractor. Building on this principle, ISCD passes teacher and student reconstructions through the same frozen HR-to-LR inverse network, isomorphic to the student. Its multi-layer features **naturally provide a matched space** for contrastive alignment **without accessing teacher hidden states or introducing cross-architecture adapters**. The inverse reconstruction term draws on cycle-consistency principles (Zhu et al., 2017) to retain agreement with the LR input.

Preliminaries and formulation. Let $\mathcal{T}$ and $\mathcal{S}$ denote the frozen teacher and trainable student. Given an LR input x, they produce HR reconstructions $\mathcal{T}(x)$ and $\mathcal{S}(x)$. ISCD employs a frozen inverse network $\mathcal{I} : \mathbb{R}^{sH \times sW \times 3} \to \mathbb{R}^{H \times W \times 3}$, whose backbone is architecturally isomorphic to the student. Both reconstructions are processed by the same $\mathcal{I}$, allowing their representations to be compared at corresponding inverse-network layers without accessing teacher hidden states.

Inverse self-consistency loss. For a mini-batch of B samples, let $\mathbf{f}^{s,(l)}_i = \mathrm{vec}(\mathcal{I}_l(\mathcal{S}(x_i)))$ and $\mathbf{f}^{t,(l)}_i = \mathrm{vec}(\mathcal{I}_l(\mathcal{T}(x_i)))$ denote the student and teacher representations at inverse layer l. After ℓ_2 normalization, the matched pair (i, i) is treated as positive, while (i, j) with $j \neq i$ provides in-batch negatives. The multi-layer objective is:

$$\mathcal{L}_{\mathrm{isc}} = \sum_{l \in \mathbb{L}} w_l \left[-\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\Big(\kappa (\hat{\mathbf{f}}^{s,(l)}_i)^\top \hat{\mathbf{f}}^{t,(l)}_i\Big)}{\sum_{j=1}^{B} \exp\Big(\kappa (\hat{\mathbf{f}}^{s,(l)}_i)^\top \hat{\mathbf{f}}^{t,(l)}_j\Big)} \right], \quad (6)$$

where $\kappa = 0.2$ is the logit scale and w_l increases with inverse-network depth. Further details are provided in A.1.5 and A.1.6, and the multi-layer design is evaluated in A.15.

Inverse reconstruction loss. We constrain the student reconstruction to remain consistent with the observed LR input, which penalizes HR reconstructions that violate the learned inverse relation:

$$\mathcal{L}_{\text{irc}} = \|\mathcal{I}(\mathcal{S}(x)) - x\|_1. \tag{7}$$

Output distillation loss. We additionally retain standard output-level supervision:

$$\mathcal{L}_{\text{out}} = \|\mathcal{S}(x) - \mathcal{T}(x)\|_1. \tag{8}$$

The **complete ISCD objective** is $\mathcal{L}_{\text{ISCD}} = \lambda_{\text{isc}}\mathcal{L}_{\text{isc}} + \lambda_{\text{irc}}\mathcal{L}_{\text{irc}} + \lambda_{\text{out}}\mathcal{L}_{\text{out}}$. The three losses occupy different numerical ranges, so λ_{out}, λ_{irc}, and λ_{isc} are chosen once per teacher to keep the weighted contributions comparable. The same values are shared across $2\times$, $3\times$, and $4\times$, remain fixed during training, and are listed in Table 6. Component, proxy, and alignment controls are provided in A.13, A.14, A.15, and A.16.

3.3 Total loss

We include a standard reconstruction loss $\mathcal{L}_{\text{REC}} = \|\mathcal{S}(x) - y_{gt}\|_1$ to preserve the student's super-resolution capability. The total loss sums $\mathcal{L}_{\text{REC}}$, $\mathcal{L}_{\text{GAMD}}$, and $\mathcal{L}_{\text{ISCD}}$:

$$\mathcal{L}_{\textbf{total}} = \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{GAMD}} + \mathcal{L}_{\text{ISCD}}. \tag{9}$$

4 Experiments

We first test whether AMID improves heterogeneous SR distillation without teacher hidden features. We then isolate what GAMD and ISCD each contribute. We finally check that the proposed mechanisms produce those gains. We fix the student to RRDB (Wang et al., 2018), a 16.7M CNN SR backbone, so that gains come from the teacher and the distillation losses, not from changing the deployed architecture. RRDB is not uniformly smaller than every teacher. $3\times$ / $4\times$ SwinIR has fewer parameters (A.2) and is used as a cross-architecture stress test, not as a compression pair. Teachers span Transformer (SwinIR (Liang et al., 2021), HAT (Chen et al., 2025)), state-space (MambaIRV2 (Guo et al., 2025)), hybrid (SRFormerV2 (Zhou et al., 2024b)), and diffusion (HYPIR (Lin et al., 2025)) families. Baselines are KD (Hinton et al., 2015), CSD (Wang et al., 2021b), AugKD (Zhang et al., 2025), and DCKD (Zhou et al., 2024a). We train on DF2K, combining 800 DIV2K-train (Agustsson & Timofte, 2017) and 2,650 Flickr2K (Lim et al., 2017b) images, for 250K iterations with batch size 8 and Adam (Kingma & Ba, 2015) at 1×10^{-4}. Data preprocessing is detailed in A.3. Model complexity is reported in A.2. Evaluation covers Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2010), Urban100 (Huang et al., 2015), BSD100 (Martin et al., 2001), and DIV2K-Val (Agustsson & Timofte, 2017). Experiments are conducted on hardware with 80 GB of dedicated memory.

4.1 Quantitative results

We report PSNR and SSIM (Wang et al., 2004) for fidelity, LPIPS (Zhang et al., 2018a) for full-reference perception, and BRISQUE (Mittal et al., 2012a), NIQE (Mittal et al., 2012b), CLIP-IQA (Wang et al., 2023), MUSIQ (Ke et al., 2021), and MANIQA (Yang et al., 2022) as no-reference scores. All metrics are computed in RGB without border cropping. Conventional Y-channel evaluation with scale-dependent cropping increases PSNR by 1.44–2.16 dB and reproduces the published SRFormerV2 results (A.5). Under that protocol, AMID attains the highest PSNR and SSIM among SwinIR-distilled students at $2\times$ ($+0.16$ dB over DCKD) and $3\times$ ($+0.07$ dB), and differs from DCKD by 0.001 dB at $4\times$ (A.6). The ranking is therefore not an artifact of RGB evaluation.

Table 2 compares AMID with output-level distillation and recent heterogeneous SR methods. KD improves fidelity only modestly, whereas CSD and AugKD remain close to, and occasionally below, the undistilled student. At $3\times$ and $4\times$, several CSD variants differ from the baseline by at most 0.03 dB, indicating that HR output matching is insufficient across heterogeneous architectures. AMID improves on most teacher–scale pairs, with a scale-dependent distribution of gains. At $2\times$, both fidelity and perception improve. SwinIR-ours is strongest on all eight metrics and exceeds DCKD by 0.16 dB PSNR. HAT-ours is strongest on seven metrics, while DCKD remains slightly higher on CLIPIQA. At $3\times$, fidelity gaps shrink and perceptual gaps increase. SwinIR-ours is again strongest on all eight metrics and raises MUSIQ from 63.32 (DCKD) to 64.54 ($+1.22$). MambaIRV2-ours is superior on six metrics, and DCKD on BRISQUE and NIQE. At $4\times$, the two methods are closer in fidelity. SwinIR-DCKD is higher by 0.016 dB PSNR, while AMID is superior on the six perceptual metrics. For SRFormerV2, AMID is superior on PSNR, SSIM, LPIPS, BRISQUE, and MANIQA, and DCKD on NIQE, CLIPIQA, and MUSIQ.

The remaining high-fidelity teachers follow this perceptual pattern (A.7). Relative to KD, HAT-ours and SRFormerV2-ours increase $3\times$ MUSIQ by 2.40 and 2.47, and MambaIRV2-ours increases $4\times$ MUSIQ from 52.16 to 58.13. HYPIR is a low-fidelity teacher (25.68 dB PSNR at $2\times$). KD declines to 30.85 dB relative to the 31.37 dB baseline, whereas AMID reaches 31.95 dB. At $4\times$, CSD attains a slightly higher PSNR (25.68 versus 25.66), while AMID is superior on the remaining metrics, including MUSIQ (48.95 versus 44.09). Because $\mathcal{L}_{\mathrm{REC}}$ is retained in all runs, the student is not trained to copy HYPIR pixels.

Table 2: Mean performance on five standard benchmarks. "baseline": student without distillation. "-ours": GAMD + ISCD. "-KD", "-CSD", "AugKD", "DCKD": KD (Hinton et al., 2015), CSD (Wang et al., 2021b), AugKD (Zhang et al., 2025), DCKD (Zhou et al., 2024a) distillation. No suffix: teacher reference. ↑/↓: higher/lower better. **Bold**: best performance among all distilled students.

Scale	Configuration	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
	baseline	31.3700	0.9142	0.0996	31.0004	5.2242	0.6203	66.8814	0.4225
	SwinIR	33.3326	0.9326	0.0743	28.0491	4.9642	0.6590	67.9030	0.4598
	SwinIR-KD	31.8943	0.9204	0.0892	30.0374	5.0793	0.6359	67.4391	0.4381
	SwinIR-CSD	31.3787	0.9142	0.0965	31.1718	5.1285	0.6109	67.2268	0.4264
	SwinIR-AugKD	31.3168	0.9137	0.0943	30.4917	5.1705	0.6145	67.5855	0.4307
	SwinIR-DCKD	32.0491	0.9225	0.0873	29.8458	5.1027	0.6370	67.6284	0.4436
2×	SwinIR-ours	**32.2072**	**0.9240**	**0.0860**	**29.6972**	**5.0277**	**0.6375**	**67.6334**	**0.4450**
	HAT	33.5723	0.9365	0.0693	27.4386	4.8866	0.6605	68.0067	0.4679
	HAT-KD	31.8936	0.9204	0.0913	30.0189	5.1038	0.6367	67.3207	0.4354
	HAT-CSD	31.3265	0.9137	0.0971	30.7002	5.1609	0.6145	67.1908	0.4260
	HAT-AugKD	31.3547	0.9141	0.0986	31.1175	5.2432	0.6189	66.9018	0.4231
	HAT-DCKD	32.0802	0.9228	0.0872	29.8120	5.0702	**0.6399**	67.6093	0.4427
	HAT-ours	**32.1713**	**0.9236**	**0.0857**	**29.4593**	**5.0182**	0.6390	**67.6854**	**0.4436**
	baseline	27.6764	0.8170	0.2493	44.5242	6.7100	0.4994	54.8755	0.3314
	SwinIR	28.7702	0.8541	0.1750	38.1315	5.4841	0.5853	65.8554	0.4382
	SwinIR-KD	28.1735	0.8380	0.2066	41.5602	5.9147	0.5471	61.0507	0.3787
	SwinIR-CSD	27.7052	0.8181	0.2454	44.3535	6.6964	0.5020	55.5315	0.3336
	SwinIR-AugKD	27.5133	0.8145	0.2374	41.4020	6.4193	0.5079	57.5598	0.3464
	SwinIR-DCKD	28.1804	0.8426	0.1939	39.7726	5.5991	0.5642	63.3150	0.4053
3×	SwinIR-ours	**28.2295**	**0.8453**	**0.1875**	**39.4604**	**5.5506**	**0.5712**	**64.5399**	**0.4206**
	MambaIRV2	29.1045	0.8589	0.1717	38.1040	5.4578	0.5923	65.8618	0.4386
	MambaIRV2-KD	28.1089	0.8367	0.2077	41.8318	5.9650	0.5492	61.0356	0.3770
	MambaIRV2-CSD	27.6700	0.8184	0.2449	44.4477	6.8197	0.5026	55.3594	0.3340
	MambaIRV2-AugKD	27.7249	0.8193	0.2416	43.9042	6.5709	0.5055	56.2361	0.3386
	MambaIRV2-DCKD	28.1735	0.8414	0.1975	**40.1020**	**5.7002**	0.5677	62.7286	0.3979
	MambaIRV2-ours	**28.1765**	**0.8428**	**0.1960**	41.1389	5.7717	**0.5686**	**63.1765**	**0.4018**
	baseline	25.5859	0.7408	0.3497	51.3503	8.0077	0.4265	44.9011	0.2767
	SwinIR	26.7267	0.7908	0.2528	43.9823	5.9344	0.5370	62.8920	0.4186
	SwinIR-KD	26.2505	0.7775	0.2855	47.4873	6.5314	0.4865	55.1865	0.3411
	SwinIR-CSD	25.6747	0.7423	0.3526	51.6350	7.6079	0.4221	44.3414	0.2734
	SwinIR-AugKD	25.6166	0.7445	0.3333	49.9826	7.2523	0.4316	48.4710	0.2942
	SwinIR-DCKD	**26.3495**	**0.7856**	0.2695	45.3454	6.0140	0.5147	59.7266	0.3803
4×	SwinIR-ours	26.3335	0.7855	**0.2669**	**45.0321**	**5.9207**	**0.5174**	**60.2025**	**0.3870**
	SRFormerV2	26.8391	0.7965	0.2451	43.4297	5.8928	0.5476	63.1789	0.4248
	SRFormerV2-KD	26.2167	0.7762	0.2870	47.5834	6.6655	0.4820	54.3117	0.3343
	SRFormerV2-CSD	25.6589	0.7415	0.3554	51.5138	7.4051	0.4238	44.1891	0.2731
	SRFormerV2-AugKD	25.6608	0.7449	0.3340	49.5544	7.3217	0.4288	47.4098	0.2913
	SRFormerV2-DCKD	26.1434	0.7824	0.2650	45.3948	**5.8368**	**0.5281**	**60.6360**	0.3846
	SRFormerV2-ours	**26.2561**	**0.7847**	**0.2648**	**44.5573**	5.8915	0.5255	60.4496	**0.3883**

Generalization to an EDSR student architecture. To examine whether AMID depends on RRDB, we replace it with EDSR (Lim et al., 2017a) and distill from SwinIR at $4\times$ (A.8). The inverse network is rebuilt to remain architecturally matched to this student. AMID is best on seven of eight metrics and outperforms DCKD on every metric, including a 0.11 dB PSNR gain and a 0.73 MUSIQ gain. KD retains a 0.09 dB PSNR advantage. Because standard EDSR is not lightweight, the comparison evaluates transfer across student backbones and is not a compression setting.

4.2 Ablation Study

The subsequent ablations examine the design of GAMD and ISCD, focusing on their complementary roles, robustness across teacher architectures, and the proposed alignment mechanisms.

Complementary roles of GAMD and ISCD. Table 3 and A.9 isolate GAMD-only (w/G), ISCD-only (w/I), and their combination. ISCD accounts for the dominant perceptual improvement, whereas GAMD provides a weaker fidelity regularizer. At $2\times$, GAMD increases PSNR by 0.10-0.27 dB, and HAT improves from 31.37 to 31.64. At $3\times$ and $4\times$, GAMD mainly prevents LPIPS and BRISQUE from deteriorating. At $3\times$, MambaIRV2-w/I raises MUSIQ from 54.88 to 62.43, while GAMD-only remains near the baseline. The combination is best on most metrics. With HAT at $4\times$, w/I is stronger on six perceptual metrics and the combination leads on PSNR and SSIM. With HYPIR at $3\times$, w/G leads on all perceptual metrics. GAMD-only still exceeds the undistilled $2\times$ student (31.62 versus 31.37 dB) despite the low teacher PSNR.

Table 3: Mean performance on five benchmark datasets. "baseline": student without distillation. Suffixes: "w/G" = distillation with GAMD only, "w/I" = ISCD only, "ours" = GAMD + ISCD. ↑/↓: higher/lower better. **Bold** indicates the best performance.

Scale	Configuration	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
2×	baseline	31.3700	0.9142	0.0996	31.0004	5.2242	0.6203	66.8814	0.4225
	HAT-w/G	31.6368	0.9175	0.0944	30.2326	5.1366	0.6295	66.9866	0.4277
	HAT-w/I	31.9655	0.9212	0.0877	29.8218	5.0656	0.6351	67.3883	0.4379
	HAT-ours	**32.1713**	**0.9236**	**0.0857**	**29.4593**	**5.0182**	**0.6390**	**67.6854**	**0.4436**
3×	baseline	27.6764	0.8170	0.2493	44.5242	6.7100	0.4994	54.8755	0.3314
	MambaIRV2-w/G	27.6820	0.8175	0.2469	44.2055	6.5864	0.5006	55.4773	0.3335
	MambaIRV2-w/I	28.1173	0.8400	0.2011	41.6866	5.8166	0.5662	62.4254	0.3933
	MambaIRV2-ours	**28.1765**	**0.8428**	**0.1960**	**41.1389**	**5.7717**	**0.5686**	**63.1765**	**0.4018**
4×	baseline	25.5859	0.7408	0.3497	51.3503	8.0077	0.4265	44.9011	0.2767
	SRFormerV2-w/G	25.6020	0.7406	0.3486	51.0282	7.4645	0.4218	45.0676	0.2787
	SRFormerV2-w/I	26.2173	0.7826	0.2697	45.3595	5.9526	0.5211	59.7020	0.3778
	SRFormerV2-ours	**26.2561**	**0.7847**	**0.2648**	**44.5573**	**5.8915**	**0.5255**	**60.4496**	**0.3883**

Effect of tail-selective attribution guidance. The inverted-Gaussian map $\mathbf{U}$ determines where attribution matching is applied. Under a HAT teacher and a GAMD-only protocol (A.10), this map attains the best PSNR, SSIM, LPIPS, MUSIQ, and MANIQA at all three scales. Uniform guidance falls below the baseline at $3\times$ (27.61 versus 27.68 dB), and monotonic guidance falls below the baseline at $4\times$ (25.46 versus 25.59 dB). Only the tail-selective map improves over the undistilled student at every scale, indicating that the U-shaped weighting is an integral component of GAMD.

GAMD requires sample-specific and spatially structured teacher attribution. Cross-sample and spatial shuffling preserve the value distribution of teacher attributions while removing sample-level and spatial correspondence (Table 15). Only the matched target improves PSNR and SSIM at all three scales, with a $2\times$ PSNR gain of 0.27 dB. Both shuffled controls remain at or below the undistilled baseline and underperform matched GAMD on $21/24$ and $20/24$ metric-scale pairs. These controls indicate that GAMD depends on the correct attribution target. Matching a shuffled map with the same value statistics is insufficient.

Attribution maps indicate selective expansion of LR support. Figs. 8, 9, and 10 provide qualitative evidence of the induced input-utilization pattern. Relative to the undistilled student, GAMD attributes a query region to an informative LR neighborhood. The additional support is concentrated on structurally similar patterns and is not uniformly distributed over the enlarged window.

Contrastive inverse alignment accounts for the teacher-dependent gains of ISCD. A loss ablation without GAMD isolates each ISCD term for HYPIR at $2\times$ (Table 17). $\mathcal{L}_{\text{out}}$ alone yields 30.85 dB and falls below the baseline. Adding $\mathcal{L}_{\text{irc}}$ reaches only 31.09 dB. $\mathcal{L}_{\text{out}} + \mathcal{L}_{\text{isc}}$ is the first configuration above the baseline at 31.72 dB. Full ISCD reaches 31.77 dB. Thus $\mathcal{L}_{\text{isc}}$ provides the teacher-dependent supervision, and $\mathcal{L}_{\text{irc}}$ acts as a complementary regularizer. The same accumulation remains beneficial for HAT, for which output distillation is already strong (Table 16).

A learned student-isomorphic inverse yields a stronger proxy than the evaluated alternatives. With SwinIR at $3\times$, the proposed inverse network leads on all eight metrics relative to a randomly initialized isomorphic proxy, a degradation-pretrained EDSR inverse, and an ImageNet-pretrained VGG extractor (Table 18). MUSIQ scores are 63.60, 61.01, 62.52, and 61.89. Across five teachers and three scales, ISCD outperforms VGG on $114/120$ comparisons (Table 20 and Fig. 11). For HYPIR at $2\times$, VGG alignment yields 30.90 dB and falls below the baseline, whereas ISCD reaches 31.77 dB. ISCD-distilled students also exhibit a narrower performance range than their teachers. At $2\times$, student PSNR spans 0.20 dB, compared with 7.99 dB among the teachers (Fig. 12).

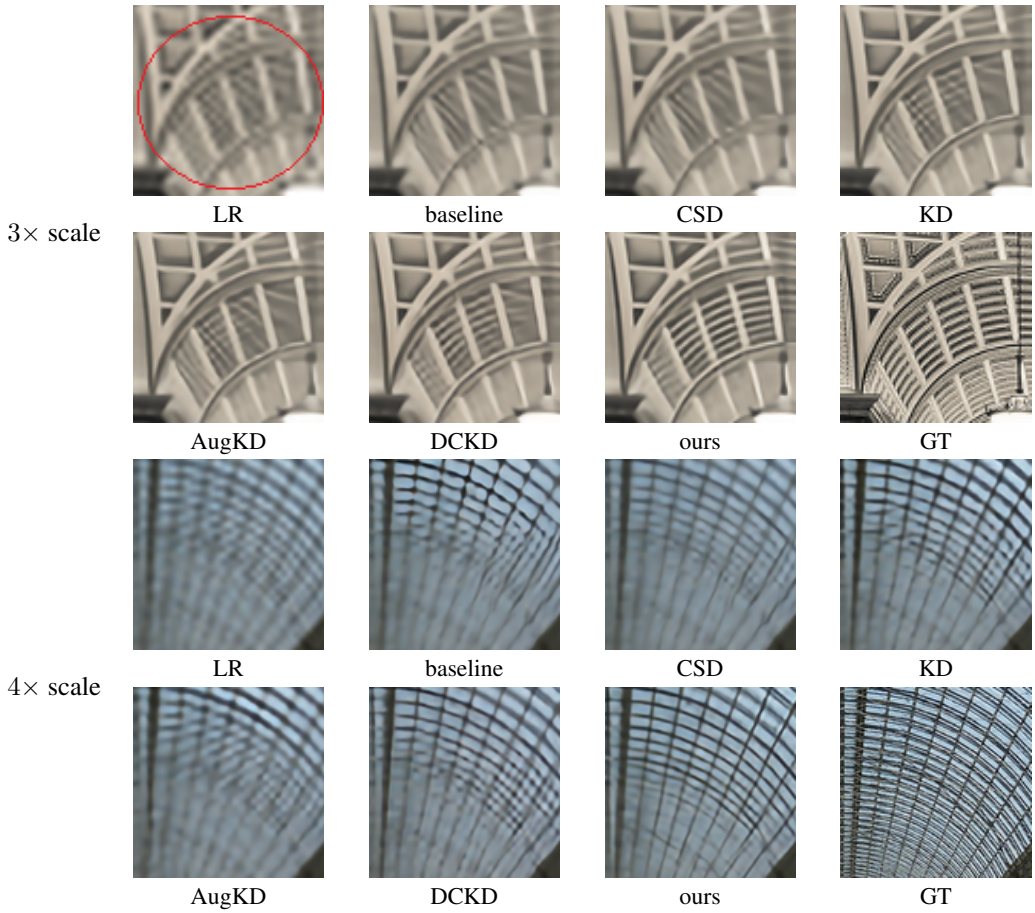


Figure 4: Qualitative comparison on representative $3\times$, and $4\times$ patches. **LR**: low-resolution input; **GT**: ground-truth. **baseline**: undistilled student; Columns labeled **CSD** (Wang et al., 2021b), **KD** (Hinton et al., 2015), **AugKD** (Zhang et al., 2025), and **DCKD** (Zhou et al., 2024a) show the student outputs distilled with the corresponding methods. **ours**: GAMD + ISCD. Fine-grained differences are best viewed when zoomed in.

Multi-layer inverse features supply complementary supervision. Under a fixed contrastive weight at $2\times$, multi-layer ISCD outperforms alignment at the deepest inverse layer(Table 19). PSNR increases from 31.91 to 31.97 dB, and LPIPS decreases from 0.0916 to 0.0866.

4.3 Qualitative results

Fig. 4 presents representative $3\times$ and $4\times$ crops with dense structure. The corresponding $2\times$ comparisons appear in Fig. 13. Additional $3\times$ and $4\times$ examples are shown in Fig. 14 and Fig. 15. AMID reconstructs sharper edges and more regular repeating textures than KD, CSD, and AugKD, and reduces residual blur near object boundaries relative to DCKD. These observations are consistent with the quantitative separation above. The same trend is observed at $2\times$ in Fig. 13.

5 Conclusion

We presented AMID, a heterogeneous SR distillation framework that uses teacher outputs and, for GAMD, teacher gradients, without reading teacher hidden features. Attribution matching and a frozen inverse cycle regularizer together improve over KD and CSD across five teachers and five benchmarks, especially on perceptual metrics at $3\times$ and $4\times$. Extending this recipe beyond super-resolution is left to future work. Limitations, including training-time attribution cost and the training-pair assumption of the ISCD inverse, are discussed in A.18.

AI USE STATEMENT

In this work, we used generative AI tools for English language polishing, grammar correction, and manuscript text refinement only. We have not used generative AI tools for research idea generation, method design, experimental implementation, data analysis, result discussion, or code writing and development, and dataset construction, model development, and graphical artifact generation are not applicable to this work. Additionally, we used generative AI tools for no recommended disclosure tasks. We have reviewed all AI-assisted work. Specifically, all AI-polished text content was fully manually checked, revised, and verified by the authors to ensure academic rigor, logical accuracy, and zero plagiarism. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Positive societal impacts. AMID enables a fixed student architecture to inherit reconstruction behavior from heterogeneous teachers during training, while leaving the deployed inference cost unchanged. This facilitates high-quality image restoration on resource-limited devices, benefiting mobile photography, medical imaging, and assistive technologies for visually impaired users.

Negative societal impacts. High-fidelity super-resolution could be misused to enhance illegally captured surveillance footage, compromising individual privacy. When combined with generative models, the technique may amplify the realism of manipulated media. However, as our method focuses on distilling existing models rather than training from scratch, it reduces the computational barrier for misuse only marginally relative to the pretrained teachers themselves. Mitigation strategies such as gated model release and misuse-monitoring mechanisms could be adopted where deployment contexts warrant such safeguards.

Our work is foundational research in heterogeneous distillation and is not tied to particular high-risk applications, but we acknowledge that advances in restoration quality can be dual-use.

REPRODUCIBILITY STATEMENT

The main text specifies the student and teacher architectures, the training schedule, the evaluation metrics, and the benchmark protocol shared by all experiments. Exact numerical replication further depends on implementation constants that are collected in Appendix A.3 and Appendix A.4. These constants include the official paired HR-LR images used as training inputs, inverse-network pretraining, the GAMD integration path and update period, the exported inverse layers, and the ISCD coefficients of each teacher. All experiments use publicly available datasets under a shared optimization budget and a fixed evaluation protocol. Ablation studies reuse the student, data, and optimizer of the corresponding main run. Source code will be released upon formal acceptance. The documented protocol together with the released implementation is intended to support independent reproduction of the reported results.

REFERENCES

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*, pp. 126–135, 2017.

Titas Anciukevicius, Patrick Fox-Roberts, Edward Rosten, and Paul Henderson. Unsupervised causal generative understanding of images. *Advances in Neural Information Processing Systems*, 35:37037–37054, 2022.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. *BMVA press*, 2012.

Qizhou Chen, Taolin Zhang, Chengyu Wang, Xiaofeng He, Dakan Wang, and Tingting Liu. Attribution analysis meets model editing: Advancing knowledge correction in vision language

models with visedit. In *AAAI Conference on Artificial Intelligence*, 2024. URL https://api.semanticscholar.org/CorpusID:271902971.

Xiangyu Chen, Xintao Wang, Wenlong Zhang, Xiangtao Kong, Yu Qiao, Jiantao Zhou, and Chao Dong. Hat: Hybrid attention transformer for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Riccardo Crupi, Alessandro Castelnovo, Daniele Regoli, and Beatriz San Miguel Gonzalez. Counterfactual explanations as interventions in latent space. *Data Mining and Knowledge Discovery*, 38(5):2733–2769, 2024.

Hangxiang Fang, Yongwen Long, Xinyi Hu, Yangtao Ou, Yuanjia Huang, and Haoji Hu. Dual cross knowledge distillation for image super-resolution. *Journal of Visual Communication and Image Representation*, 95:103858, 2023. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2023.103858. URL https://www.sciencedirect.com/science/article/pii/S1047320323001086.

Song Fei, Tian Ye, Lujia Wang, and Lei Zhu. Lucidflux: Caption-free universal image restoration via a large-scale diffusion transformer. In *International Conference on Learning Representations*, 2025.

Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. *CoRR*, abs/2011.11036, 2020. URL https://arxiv.org/abs/2011.11036.

Guangyu Guo, Dingwen Zhang, Longfei Han, Nian Liu, Ming-Ming Cheng, and Junwei Han. Pixel distillation: Cost-flexible distillation across image sizes and heterogeneous networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9536–9550, 2024a. doi: 10.1109/TPAMI.2024.3421277.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision*, pp. 222–241. Springer, 2024b.

Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28124–28133, 2025.

Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2154–2164, 2021.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems*, 36:79570–79582, 2023.

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)*, pp. 518–522, 2020. doi: 10.1109/ICIP40778.2020.9190917.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://dblp.uni-trier.de/db/journals/corr/corr1503.html#HintonVD15.

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*, pp. 5197–5206, 2015.

Yuxuan Jiang, Chen Feng, Fan Zhang, and David Bull. *MTKD: Multi-Teacher Knowledge Distillation for Image Super-Resolution*, pp. 364–382. Springer Nature Switzerland, October 2024. ISBN 9783031729331. doi: 10.1007/978-3-031-72933-1_21. URL http://dx.doi.org/10.1007/978-3-031-72933-1_21.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Guopeng Li, Qiang Wang, Ke Yan, Shouhong Ding, Yuan Gao, and Gui-Song Xia. Fuse before transfer: Knowledge fusion for heterogeneous distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3445–3454, 2025a.

Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. In *ICML*, 2025b.

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*, abs/2404.02573, 2024. URL https://api.semanticscholar.org/CorpusID:268876310.

Xiaohui Li, Yihao Liu, Shuo Cao, Ziyan Chen, Shaobin Zhuang, Xiangyu Chen, Yinan He, Yi Wang, and Yu Qiao. Diffvsr: Revealing an effective recipe for taming robust video super-resolution against complex degradations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15319–15328, 2025c.

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*, pp. 1833–1844, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. *CoRR*, abs/1707.02921, 2017a. URL http://arxiv.org/abs/1707.02921.

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*, pp. 136–144, 2017b.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2017. URL https://api.semanticscholar.org/CorpusID:206771220.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM Transactions on Graphics (TOG)*, 44(6):1–21, 2025.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International conference on learning representations*, 2021.

Zhengyang Lu, Bingjie Lu, and Feng Wang. Causalsr: Structural causal model-driven super-resolution with counterfactual inference. *Neurocomputing*, 646:130375, 2025.

Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8046–8056, 2022.

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*, volume 2, pp. 416–423. Ieee, 2001.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. *CoRR*, abs/1902.03393, 2019. URL http://arxiv.org/abs/1902.03393.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012b.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. *CoRR*, abs/1904.05068, 2019. URL http://arxiv.org/abs/1904.05068.

Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. ALP-KD: attention-based layer projection for knowledge distillation. *CoRR*, abs/2012.14022, 2020. URL https://arxiv.org/abs/2012.14022.

Ruihong Qiu, Sen Wang, Zhi Chen, Hongzhi Yin, and Zi Huang. Causalrec: Causal inference for visual debiasing in visually-aware recommendation. *CoRR*, abs/2107.02390, 2021. URL https://arxiv.org/abs/2107.02390.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets. *arXiv e-prints*, art. arXiv:1412.6550, December 2014. doi: 10.48550/arXiv.1412.6550.

Suho Son, Jeonghyeok Park, and Byung Cheol Song. Mask-based adaptive response distillation for efficient image super-resolution. *Neurocomputing*, 669:132533, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2025.132533. URL https://www.sciencedirect.com/science/article/pii/S0925231225032059.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. URL http://arxiv.org/abs/1703.01365.

Vimanyu Taneja and Soumya Banerjee. RIGHT ANSWERS, WRONG REASONS: DISSOCIATING UNDERSTANDING FROM CORRECTNESS IN LLM REASONING. In *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*, 2026. URL https://openreview.net/forum?id=LndEvcJ5Rt.

Huiyuan Tian, Bonan Xu, Shijian Li, and Xin Jin. From low-rank features to encoding mismatch: Rethinking feature distillation in vision transformers. *arXiv preprint arXiv:2511.15572*, 2025.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *CoRR*, abs/1910.10699, 2019. URL http://arxiv.org/abs/1910.10699.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. 2024.

Penghao Wang, Yuhao Zhou, Mengxuan Wu, Panpan Zhang, Zhangyang Wang, and Kai Wang. Data Efficient Any Transformer-to-Mamba Distillation via Attention Bridge. *arXiv e-prints*, art. arXiv:2510.19266, October 2025. doi: 10.48550/arXiv.2510.19266.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. *CoRR*, abs/2108.08782, 2021a. URL https://arxiv.org/abs/2108.08782.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018.

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. *CoRR*, abs/2105.11683, 2021b. URL https://arxiv.org/abs/2105.11683.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Hongjun Wu, Li Xiao, Xingkuo Zhang, and Yining Miao. Aligning in a compact space: Contrastive knowledge distillation between heterogeneous architectures. *arXiv preprint arXiv:2405.18524*, 2024.

Wanxin Xiao, Yafei Zhang, Hongbin Wang, Fan Li, and Hua Jin. Heterogeneous knowledge distillation for simultaneous infrared-visible image fusion and super-resolution. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022. doi: 10.1109/TIM.2022.3149101.

Jiao Xie, Linrui Gong, Shitong Shao, Shaohui Lin, and Linkai Luo. Hybrid knowledge distillation from intermediate layers for efficient single image super-resolution. *Neurocomputing*, 554:126592, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.126592. URL https://www.sciencedirect.com/science/article/pii/S0925231223007154.

Dizhan Xue, Shengsheng Qian, and Changsheng Xu. Variational causal inference network for explanatory visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2515–2525, 2023.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1191–1200, 2022.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9847–9857, 2021.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Yuhui Wu, and Lei Zhang. Fine-structure preserved real-world image super-resolution via transfer vae training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12415–12426, 2025.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7130–7138, 2017. doi: 10.1109/CVPR.2017.754.

Min Yoon, Seunghyun Lee, and Byung Cheol Song. Takdsr: Teacher assistant knowledge distillation framework for graphics image super-resolution. *IEEE Access*, 11:112015–112026, 2023. doi: 10.1109/ACCESS.2023.3323273.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25669–25680, 2024.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *CoRR*, abs/1612.03928, 2016. URL http://arxiv.org/abs/1612.03928.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pp. 711–730. Springer, 2010.

Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in neural information processing systems*, 33:655–666, 2020.

Kangkai Zhang, Shiming Ge, Ruixin Shi, and Dan Zeng. Low-resolution object recognition with cross-resolution relational contrastive distillation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2374–2384, 2024a. doi: 10.1109/TCSVT.2023.3310042.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018a.

Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European conference on computer vision*, pp. 649–667. Springer, 2022.

Yiman Zhang, Hanting Chen, Xinghao Chen, Yiping Deng, Chunjing Xu, and Yunhe Wang. Data-free knowledge distillation for image super-resolution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7848–7857, 2021. doi: 10.1109/CVPR46437.2021.00776.

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

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. *CoRR*, abs/1802.08797, 2018c. URL http://arxiv.org/abs/1802.08797.

Yun Zhang, Wei Li, Simiao Li, Hanting Chen, Zhijun Tu, Bingyi Jing, Shaohui Lin, Jie Hu, and Wenjia Wang. Augkd: Ingenious augmentations empower knowledge distillation for image super-resolution. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhizhong Zhang, Yuan Xie, Chong Zhang, Yanbo Wang, Yanyun Qu, Shaohui Lin, Lizhuang Ma, and Qi Tian. Dynamic image super-resolution via progressive contrastive self-distillation. *Pattern Recognition*, 153:110502, 2024b.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962, 2022.

Yunshuai Zhou, Junbo Qiao, Jincheng Liao, Wei Li, Simiao Li, Jiao Xie, Yunhang Shen, Jie Hu, and Shaohui Lin. Dynamic contrastive knowledge distillation for efficient image restoration. In *AAAI Conference on Artificial Intelligence*, 2024a. URL https://api.semanticscholar.org/CorpusID:274656278.

Yupeng Zhou, Zhen Li, Chun-Le Guo, Li Liu, Ming-Ming Cheng, and Qibin Hou. Srformerv2: Taking a closer look at permuted self-attention for image super-resolution, 2024b. URL https://arxiv.org/abs/2303.09735.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Wencheng Zhu, Xin Zhou, Pengfei Zhu, Yu Wang, and Qinghua Hu. Ckd: Contrastive knowledge distillation from a sample-wise perspective. *IEEE Transactions on Image Processing*, 2025.

A APPENDIX

The content of the supplementary material involves:

- AMID formulation details in A.1.
- Parameter and capacity comparison of student and teacher models in A.2.
- Training data preprocessing in A.3.
- Implementation constants and ISCD coefficients in A.4.
- Evaluation-protocol calibration in A.5.
- Standard Y-channel evaluation across scales in A.6.
- Additional teacher–scale comparisons in A.7.
- Transfer to an EDSR student architecture in A.8.
- Complementary effects of GAMD and ISCD in A.9.
- Ablation study of GAMD in A.10.
- Attribution-correspondence controls for GAMD in A.11.
- Attribution maps at different scales in A.12.
- Ablation study of ISCD loss components in A.13.
- Additional controls for the ISCD inverse proxy in A.14.
- Single-layer versus multi-layer inverse alignment in A.15.
- Student-isomorphic inverse alignment versus VGG-based feature matching in A.16.
- Additional visual results of AMID in A.17.
- Limitations in A.18.

A.1 AMID FORMULATION DETAILS

A.1.1 SUPPLEMENTARY ATTRIBUTION FORMULATION

We provide a detailed interpretation of each term in equation 1, whose formulation follows the Local Attribution Map framework (Gu & Dong, 2020). The term $\partial f_p(\gamma(t))/\partial\gamma_i(t)$ measures the sensitivity of output element p with respect to input element i at path position t, while $\dot{\gamma}_i(t) = \partial\gamma_i(t)/\partial t$ measures the instantaneous rate of change of that input element along the Gaussian-blur path. Their product therefore quantifies the instantaneous contribution rate at each path position. Integrating this product over the complete path yields $\mathrm{Att}_{p,i}$, which measures the overall contribution of input pixel i to output pixel p.

Rather than relying on the gradient at the final input x alone, the path integral over t accumulates gradient contributions along the Gaussian-blur path $\gamma(t)$ from x' to x. This accumulation accounts for network nonlinearities and yields a more faithful and stable attribution. Computing $\mathrm{Att}_{p,i}$ for every input pixel i with respect to a fixed output pixel p produces an attribution map of the same spatial resolution as the input image x, which serves as the core supervision signal in GAMD.

A.1.2 MOTIVATION FOR THE ℓ_2 NORM IN GUIDANCE MAP ESTIMATION

We elaborate on the rationale for adopting the ℓ_2 norm in both the pixel-wise residual computation and the subsequent channel-wise aggregation step of the guidance map estimation. Compared to ℓ_1 norm, ℓ_2 norm exhibits greater sensitivity to larger errors and amplifies their contribution. This characteristic is particularly advantageous for super-resolution tasks, as it aligns with the heightened perceptual sensitivity of the human visual system to high-frequency details, including edges and textures. Consequently, the ℓ_2 norm emphasizes these regions, so that the guidance map emphasizes perceptually salient residuals when forming $\hat{R}$, as evidenced by the grayscale visualizations in Figure 5.

For channel aggregation, the Euclidean norm combines the squared residuals across channels and captures the joint energy of channel-wise errors. This strategy is well-suited for color images, where



Figure 5: Spatial instance of the channel-aggregated residual $\mathbf{R}$, the batch-normalized residual $\hat{R}$, and the guidance map $\mathbf{U} = u(\hat{R})$ on one HR patch.

channel correlations are most meaningful when considered holistically. Moreover, it maintains consistency with the pixel-wise ℓ_2 residual computation while avoiding the potential error cancellation that can arise from simple mean aggregation.

A.1.3 Details of Gaussian mapping for attribution guidance

We provide the complete formulation of the Gaussian-shaped mapping that converts normalized residuals to guidance weights, complementing the description in the main text. The mapping adopts a Gaussian function with parameters $\mu = 0.5$ and $\sigma = 1/6$ as a symmetric unimodal reference shape, not as a probabilistic model. This parameterization places the interval $[\mu - 3\sigma, \mu + 3\sigma]$ exactly at $[0, 1]$, covering the full range of $\hat{R}$. For each pixel value $x_p = \hat{R}_{n,1,sh,sw}$, the Gaussian response is computed as:

$$g(x_p) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_p - 0.5)^2}{2\sigma^2}\right). \tag{10}$$

Since $g(x_p)$ attains its maximum $g_{\max} = 1/(\sqrt{2\pi}\sigma)$ at $x_p = 0.5$ and its minimum $g_{\min} = \min\{g(0), g(1)\}$ at the interval endpoints, we linearly normalize $g(x_p)$ to $[0, 1]$ to obtain a normalized mid-range response:

$$\bar{g}(x_p) = \frac{g(x_p) - g_{\min}}{g_{\max} - g_{\min}}. \tag{11}$$

The guidance weight is the complement of this mid-bulk response, $u(x_p) = 1 - \bar{g}(x_p)$, so both tails receive high weight.

$$u(x_p) = 1 - \bar{g}(x_p). \tag{12}$$

This mapping ensures that moderate residual values near 0.5 receive low guidance weight, while extreme residual values near 0 or 1 receive high guidance weight. The resulting guidance map $\mathbf{U} \in [0, 1]^{N \times 1 \times sH \times sW}$ is used to weight the attribution alignment loss as described in the main text.

A.1.4 Implementation and optimization details of GAMD

GAMD computes path-integrated attribution for the scalar response in equation 3. Let $\{\gamma_k\}_{k=1}^{K}$ denote the samples along the Gaussian-blur path and let $\dot{\gamma}_k$ denote the corresponding discrete path derivatives. The attribution tensor for model $m \in \{T, S\}$ is approximated as:

$$\Phi^m(x) \approx \frac{1}{K} \sum_{k=1}^{K} \nabla_{\gamma_k} \mathcal{R}^m(\gamma_k; \mathbf{U}) \odot \dot{\gamma}_k, \tag{13}$$

where $\odot$ denotes element-wise multiplication. We use $K = 50$ path samples. GAMD is applied every N_{GAMD} iterations and is skipped on the remaining steps. The period $N_{\text{GAMD}} \in \{50, 100\}$ is chosen once per teacher, is shared across scales, and is listed in Table 6. The same guidance map $\mathbf{U}$ is held fixed across all path samples and is applied identically to the teacher and student responses.

Because attribution magnitudes can differ across models and input channels, we normalize each channel independently:

$$\bar{\Phi}^m_{c,h,w} = \frac{\left|\Phi^m_{c,h,w}\right|}{\max_{h',w'} \left|\Phi^m_{c,h',w'}\right| + \epsilon}, \tag{14}$$

where $\epsilon > 0$ ensures numerical stability. This operation preserves the spatial distribution within each channel while reducing sensitivity to model-dependent attribution scale. The normalized teacher and student tensors are then compared using equation 5.

Since the teacher is frozen and $\mathbf{U}$ is treated as fixed during GAMD computation, optimization acts through the normalized student attribution. At points where the absolute-value and maximum operations are differentiable, the gradient is:

$$\frac{\partial \mathcal{L}_{\mathrm{GAMD}}}{\partial \theta_S} = \frac{1}{|\Omega_x|} \sum_{i \in \Omega_x} \operatorname{sign}\left(\bar{\Phi}_i^S - \bar{\Phi}_i^T\right) \frac{\partial \bar{\Phi}_i^S}{\partial \theta_S}. \tag{15}$$

where $\operatorname{sign}(\cdot)$ is the derivative of the absolute value, defined as $+1$ for positive arguments, -1 for negative arguments, and 0 at zero. The derivative $\partial \bar{\Phi}^S / \partial \theta_S$ includes the normalization derivative and the mixed derivatives induced by differentiating the input-gradient attribution with respect to the student parameters. In practice, these terms are handled by automatic differentiation. No gradient is propagated through the residual-based guidance map.

A.1.5 Choice of contrastive alignment

At each exported inverse-network layer, teacher and student features generated from the same LR input form a positive pair, while teacher features from other samples in the mini-batch serve as negatives. This structure naturally defines a B-way identification objective in which each student representation must retrieve its corresponding teacher representation.

Unlike point-wise ℓ_1 or ℓ_2 regression, which treats each teacher-student pair independently, the contrastive objective evaluates a positive match relative to other samples in the batch. It therefore preserves sample-specific structure and discourages different inputs from collapsing to similar inverse representations. This choice is particularly suitable for ISCD because both representations are produced by the same frozen inverse network and consequently share dimensionality and layer semantics. We do not assume that contrastive alignment is universally superior to regression-based alternatives.

For $L = |\mathbb{L}|$ exported layers ordered from shallow to deep, we use:

$$w_l = \frac{l}{L}, \qquad l = 1, \ldots, L, \tag{16}$$

so that deeper inverse representations receive progressively larger weights. The contrastive logits are multiplied by $\kappa = 0.2$, which is a logit scale rather than a conventional divisive temperature. The outer coefficient λ_{isc} is calibrated jointly with the output and inverse-reconstruction coefficients to balance their weighted magnitudes.

The contribution of contrastive alignment is isolated in Appendix A.13. Single- versus multi-layer alignment is evaluated in Appendix A.15, while comparisons with alternative proxy spaces are provided in Appendix A.14 and Appendix A.16.

A.1.6 Inverse-network pretraining

The inverse network is pretrained on the same official DF2K pairs used by the student, with the input and output reversed. Each pair consists of an official HR image and the LR image paired with it at the training scale. The inverse network learns the map from that HR image to the paired LR observation. We introduce no additional dataset, augmentation, or sampling strategy.

The inverse network is optimized only with an ℓ_1 reconstruction loss:

$$\mathcal{L}_{\mathrm{inv}} = \mathbb{E}_{(I_{\mathrm{HR}}, I_{\mathrm{LR}})} \left[\| G_{\mathrm{inv}}(I_{\mathrm{HR}}) - I_{\mathrm{LR}} \|_1 \right], \tag{17}$$

where G_{inv} denotes the inverse network. No perceptual, adversarial, contrastive, or feature-matching terms are used. The number of training iterations is identical to that of the LR-to-HR student SR network.

After pretraining, all parameters of G_{inv} are frozen. The inverse network serves only as a fixed feature extractor and teacher for the contrastive alignment above. This keeps its representations

determined by the student SR training data and optimization budget and prevents co-adaptation or drift during ISCD.

The inverse network keeps the residual-dense trunk of the RRDB student and replaces the upsampling head with strided downsampling modules that reduce spatial resolution by the training scale s. The resulting map is $\mathcal{I} : \mathbb{R}^{sH \times sW \times 3} \to \mathbb{R}^{H \times W \times 3}$. Residual-dense width and depth follow the student. We export L feature maps for $\mathcal{L}_{\text{isc}}$. The deepest map is taken after the second scale-conversion block and immediately before the reconstruction head that projects latent features to the three-channel LR output, matching the single-layer control in Appendix A.15. The remaining maps are taken after the preceding residual-dense groups of the trunk. Layer weights are $w_l = l/L$ for $l = 1, \ldots, L$.

A.2 Parameter and capacity comparison of student and teacher models

Table 4 reports the parameter count, FLOPs, and on-disk weight size of the RRDB student and the five teachers at $2\times$, $3\times$, and $4\times$. The teachers span Transformer, state-space, hybrid, and diffusion families, and their capacities are not uniformly larger than that of the student. RRDB contains 16.7M parameters with an on-disk weight file of about 64 MB across scales, while HYPIR remains larger than the remaining teachers by more than an order of magnitude in both parameters and FLOPs because of its diffusion inference pipeline.

Table 4: Inference complexity of the student and teachers with a shared 256×256 low-resolution input. Parameter counts and FLOPs are measured on the instantiated official models. Model size is the operating-system file size of the released weight file.

Scale	Models	Params(M)	FLOPs(G)	Model Size(MB)
$2\times$	Student-RRDB	16.703	587.597	63.9
	Tea-SwinIR(Liang et al., 2021)	11.752	1537.45	64.1
	Tea-HAT(Chen et al., 2025)	40.699	5297.19	157
	Tea-MambaIRV2(Guo et al., 2025)	34.091	4151.39	131
	Tea-SRFormerV2(Zhou et al., 2024b)	21.92	3620.83	240
	Tea-HYPIR(Lin et al., 2025)	1549.43	54316.36	989
$3\times$	Student-RRDB	16.698	2280.48	63.9
	Tea-SwinIR(Liang et al., 2021)	11.94	1562.75	64.8
	Tea-HAT(Chen et al., 2025)	40.88	5322.48	158
	Tea-MambaIRV2(Guo et al., 2025)	34.28	4176.68	132
	Tea-SRFormerV2(Zhou et al., 2024b)	22.10	3652.84	241
	Tea-HYPIR(Lin et al., 2025)	1549.43	54316.36	989
$4\times$	Student-RRDB	16.70	2349.71	63.9
	Tea-SwinIR(Liang et al., 2021)	11.90	1617.48	64.7
	Tea-HAT(Chen et al., 2025)	40.85	5377.22	158
	Tea-MambaIRV2(Guo et al., 2025)	34.24	4231.42	132
	Tea-SRFormerV2(Zhou et al., 2024b)	22.07	3722.11	241
	Tea-HYPIR(Lin et al., 2025)	1549.43	54316.36	989

Parameter counts and FLOPs are computed from the instantiated official models at runtime under a 256×256 low-resolution input and a single forward pass. Model size is the Windows file size of the released checkpoint associated with each model. For HYPIR, the runtime parameter count covers the assembled diffusion pipeline, including pretrained modules loaded from additional weight files and caches, while the listed 989 MB is the official HYPIR checkpoint only. For SRFormerV2, the official checkpoint stores EMA copies and other tensors that are excluded from the runtime parameter count and do not enter the forward pass, which makes the 240 MB file larger than the 22M counted parameters. HYPIR follows its official diffusion inference procedure, so its FLOPs exceed those of the other teachers by more than an order of magnitude and remain constant across scales.

RRDB follows the BasicSR implementation in which the $2\times$ network applies pixel unshuffle before the residual-dense trunk. A 256×256 input is therefore rearranged to 128×128, and the trunk FLOPs fall to about one quarter of the $3\times$ and $4\times$ entries, which keep the trunk at 256×256. The small gap between $3\times$ and $4\times$ RRDB comes from the extra upsampling stage at $4\times$. Window-based

teachers do not use this unshuffle and therefore keep nearly constant FLOPs across scales under the same LR input.

Teacher selection and capacity asymmetry. Table 4 summarizes parameter count, FLOPs, and on-disk weight size. The experiment studies heterogeneous distillation into a fixed deployment model and is not designed as strict large-to-small compression. Teacher intermediate features are unused. GAMD requires differentiable access to the teacher during offline training. The student is the only model deployed at inference, so teacher complexity characterizes the external supervision source and does not set the deployed runtime budget.

We include five teachers to cover a wide architectural spectrum, without enforcing a uniform parametric ordering relative to RRDB. HAT, MambaIRV2, SRFormerV2, and HYPIR are heavier cloud-side supervisors, and HYPIR exceeds the other teachers by more than an order of magnitude in parameters and FLOPs because of its diffusion inference pipeline. SwinIR is included as a cross-architecture stress test between a CNN student and a Transformer teacher, which is central to the heterogeneous setting and is not intended as a canonical larger-teacher pair.

This distinction matters at $3\times$ and $4\times$, where SwinIR has fewer parameters than RRDB (about 11.9M versus 16.7M) and lower FLOPs under the reported 256×256 input protocol, for example 1617.48G versus 2349.71G at $4\times$. These numbers do not make SwinIR a weak or inappropriate teacher. In deployment, a cloud model is often exposed only through output APIs, and its size relative to the on-device student may be unknown in advance. A teacher may therefore be smaller on one complexity axis while still providing reconstruction behavior that the student cannot obtain from ground-truth supervision alone. In our main experiments (Table 2), zero-shot SwinIR substantially outperforms the undistilled RRDB student despite the near-capacity profile, for example $2\times$ PSNR 33.33 versus 31.37 dB and $3\times$ PSNR 28.77 versus 27.68 dB. The relevant question is whether AMID transfers useful cross-architecture knowledge from such a teacher, not whether the teacher exceeds the student on every parameter or FLOP count.

The cross-scale FLOP pattern further shows that parameter count is an incomplete proxy for deployment cost in SR. RRDB and the Transformer teachers follow different compute profiles under the same input size, so a teacher that is smaller in parameters can still differ from the student in both architecture and runtime behavior. The evaluation therefore spans both clearly asymmetric pairs (HAT, MambaIRV2, SRFormerV2, HYPIR) and a near-capacity cross-architecture pair (SwinIR), under a single fixed student and training protocol. AMID is intended to transfer teacher-specific reconstruction and reasoning behavior in all of these regimes, including when the teacher is not uniformly larger than the student. The setting is heterogeneous distillation for on-device deployment, and it is not strict parametric compression in every teacher-student pair.

Training-time compute Table 4 compares inference-time parameter count, FLOPs, and on-disk weight size among the student and teachers, while the *Training Cost* column in Table 1 summarizes relative training-time overhead among distillation paradigms. These labels are qualitative and compare methods *with each other*, not the absolute cost of undistilled RRDB training. They reflect extra forward and backward passes, auxiliary networks, feature alignment, and data-generation steps incurred during distillation.

We use four coarse levels. *Very low* (KD) denotes output-level supervision with cached teacher predictions and negligible additional training compute beyond the student update itself. *Low* (CSD) projects SR outputs through a pretrained VGG extractor and applies a contrastive loss in that space, remaining inexpensive relative to multi-stage or data-driven pipelines. *Moderate* (FAKD and AMID) denotes methods whose training requires nontrivial additional computation beyond a single output loss, such as multi-component losses, periodic attribution computation, or repeated passes through auxiliary networks. *High* (MTKD and DCKD) denotes distillation pipelines with substantially heavier training budgets, including multi-teacher optimization and data-driven contrastive distillation with large auxiliary compute. AMID is placed in the *Moderate* category because its training cost exceeds lightweight output distillation, but remains well below the high-cost data-driven reference DCKD (Zhou et al., 2024a) under our profiling.

To make this concrete, we profile GFLOPs per training iteration under our main setting ($2\times$ SR, GT crop 384×384, batch size 8, 250K iterations). The undistilled baseline trains RRDB with $\mathcal{L}_{\mathrm{REC}}$ only and defines the $1.0\times$ reference. KD is approximately $1.0\times$ because teacher outputs can be

cached. AMID adds two training-only components. ISCD runs the frozen inverse network $\mathcal{I}$ on teacher and student outputs every iteration, contributing a fixed 2644.2 GFLOPs/iter regardless of teacher architecture once outputs are available.

GAMD computes path-integrated gradients every N_{GAMD} iterations, with $N_{\mathrm{GAMD}} \in \{50, 100\}$ chosen once per teacher as in Table 6. We report this cost amortized over all iterations. Table 5 and Fig. 6 list and visualize the resulting AMID totals at $2\times$. For example, with HAT, AMID requires 8601.0 GFLOPs/iter ($8.7\times$ baseline), whereas DCKD under the same teacher, crop size, and iteration budget requires $23.2\times$ baseline. With SwinIR, AMID totals 5428.8 GFLOPs/iter ($5.5\times$), and with HYPIR it reaches 8903.8 GFLOPs/iter ($9.0\times$) because GAMD cost scales with teacher forward and backward complexity. At $4\times$ with HYPIR, AMID training cost is $7.1\times$ baseline, compared with $7.8\times$ for DCKD under the same protocol.

Two points clarify the deployment implication. First, all AMID overhead is confined to training; at inference, only the 16.7 M RRDB student is deployed, with the same inference FLOPs as the baseline student in Table 4. Second, GFLOPs/iter is a hardware-agnostic compute proxy; peak memory during AMID training is dominated mainly by the frozen $\mathcal{I}$ forward pass and periodic GAMD backward passes. Overall, the middle label in Table 1 indicates that AMID trades a moderate one-time offline training budget for persistent on-device efficiency. Fig. 6 shows that AMID at $2\times$ spans $5.5\times$-$9.0\times$ baseline across teachers, whereas DCKD (HAT) requires $23.2\times$ under the same crop size and iteration budget.

Table 5: AMID training compute at $2\times$ (GT crop 384×384; GFLOPs per iteration). Baseline: RRDB with $\mathcal{L}_{\mathrm{REC}}$ only ($1.0\times$). ISCD: fixed frozen-$\mathcal{I}$ cost. GAMD: amortized over iterations (applied every 50-100 iterations). DCKD (Zhou et al., 2024a) profiled under the same crop, scale, and 250K-iteration budget with HAT.

Teacher	ISCD (G)	GAMD amort. (G)	AMID total (G)	$\times$ vs. baseline
SwinIR	2644.2	1793.0	5428.8	$5.5\times$
HAT	2644.2	4965.3	8601.0	$8.7\times$
SRFormerV2	2644.2	3550.9	7186.6	$7.2\times$
MambaIRV2	2644.2	3998.5	7634.3	$7.7\times$
HYPIR	2644.2	5268.1	8903.8	$9.0\times$
DCKD (HAT, $2\times$): $23.2\times$ baseline				

A.3 Training data preprocessing

Training uses the official DF2K split formed from 800 DIV2K training images (Agustsson & Timofte, 2017) and 2,650 Flickr2K images (Lim et al., 2017b). High-resolution supervision comes from the official HR images. Low-resolution inputs are the official LR images paired with these HR images at the target scale. Each HR image is randomly cropped to 384×384. The LR patch is taken from the matched official LR image after the crop coordinates are scaled by $1/s$. Random horizontal flipping is applied jointly to both patches.

A.4 Implementation constants and ISCD coefficients

The outer objective uses unit weights on $\mathcal{L}_{\mathrm{REC}}$, $\mathcal{L}_{\mathrm{GAMD}}$, and $\mathcal{L}_{\mathrm{ISCD}}$. Inside ISCD the raw terms occupy different numerical ranges, so λ_{out}, λ_{irc}, and λ_{isc} are chosen once per teacher to bring the three weighted contributions to a comparable scale at the start of training. The same coefficients are then used at $2\times$, $3\times$, and $4\times$ for that teacher and remain fixed. Ablations that disable one ISCD term keep the remaining coefficients unchanged.

Table 6 lists the coefficients used by the main AMID runs and by all ISCD-only controls of the same teacher. The logit scale of the contrastive term is $\kappa = 0.2$. GAMD uses $K = 50$ path samples and the update period N_{GAMD} shown in the table.

The EDSR-student study at $4\times$ with a SwinIR teacher uses the SwinIR coefficients in Table 6. Its inverse network follows the EDSR trunk and uses the same HR-to-LR head replacement as the RRDB inverse.

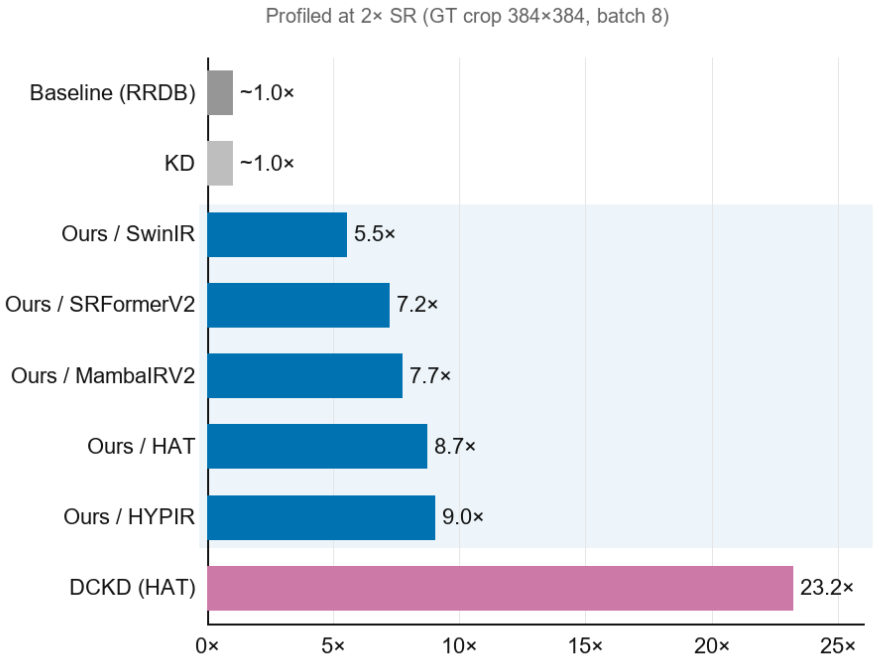


Figure 6: Relative AMID training cost at $2\times$ (GT crop 384×384, batch size 8; profiled GFLOPs per iteration, normalized to undistilled RRDB baseline $= 1.0\times$). *Ours / Teacher*: AMID totals for each heterogeneous teacher; *DCKD (HAT)*: data-driven contrastive distillation under the same crop, scale, and 250K-iteration budget (Table 5).

Table 6: ISCD coefficients shared by the $2\times$, $3\times$, and $4\times$ runs of each teacher. The corresponding ISCD-only ablations use the same values.

Teacher	λ_{out}	λ_{irc}	λ_{isc}	N_{GAMD}
SwinIR	0.1	1.0	0.01	50
HAT	0.1	1.0	0.01	50
MambaIRV2	0.1	1.0	0.05	100
SRFormerV2	0.1	1.0	0.01	100
HYPIR	0.1	1.0	0.05	100

A.5 Evaluation-protocol calibration

Our main experiments compute all metrics in RGB without border removal, whereas conventional SR benchmarks report PSNR and SSIM on the luminance channel after cropping boundary pixels according to the upscaling factor. To verify our evaluation implementation and quantify the resulting difference, we evaluate the official SRFormerV2 $4\times$ model under three protocols: RGB without border removal, Y-channel without border removal, and Y-channel after cropping four boundary pixels. We additionally compare the final protocol with the values published by SRFormerV2 (Zhou et al., 2024b). Our main experiments evaluate five benchmarks, Set5, Set14, BSD100, Urban100, and DIV2K-Val. The published SRFormerV2 results cover the first four datasets but do not report DIV2K-Val. We therefore perform protocol calibration on the four shared benchmarks.

As shown in Table 7, the reproduced Y+crop results match the published SRFormerV2 values on all four benchmarks, validating our metric implementation. Converting RGB evaluation to the Y channel increases PSNR by 1.34-1.91 dB, while border cropping contributes a further 0.10-0.34 dB. Together, the standard protocol increases PSNR by 1.44-2.16 dB. The lower absolute PSNR values in our main tables therefore arise from the stricter RGB full-image protocol rather than from an inconsistency in model reproduction.

Table 7: Evaluation-protocol calibration on the four benchmarks shared with the official SRFormerV2 $4\times$ model. Each entry reports PSNR/SSIM. **RGB** (evaluation protocol used in our main experiment) and **Y** use the full image; **Y+crop** removes four boundary pixels. **Ref** refers to results which are published in SRFormerV2 (Zhou et al., 2024b). Δ_{color} denotes Y$-$RGB, and Δ_{standard} denotes (Y+crop)$-$RGB.

Dataset	RGB reproduced	Y only	Y+crop	**Ref**	Δ_{color}	Δ_{standard}
Set5	30.9093/0.8937	32.8190/0.9061	33.0669/0.9066	33.06/0.9066	1.9097/0.0124	2.1576/0.0129
Set14	27.0752/0.7802	28.8734/0.7984	29.2130/0.7978	29.21/0.7978	1.7982/0.0182	2.1378/0.0176
BSD100	26.5395/0.7341	27.8840/0.7528	27.9806/0.7522	27.98/0.7522	1.3445/0.0187	1.4411/0.0181
Urban100	26.3157/0.8277	27.8713/0.8391	28.0372/0.8391	28.04/0.8391	1.5556/0.0114	1.7215/0.0114

A.6 Standard Y-channel evaluation across scales

The protocol calibration above verifies that our evaluation implementation reproduces the published SRFormerV2 results. We further examine whether the relative ranking of the compared methods is preserved under the conventional SR protocol. For each scale s, predictions are converted to the Y channel and s boundary pixels are removed before computing PSNR and SSIM. Table 8 reports mean performance over all five benchmarks used in our main evaluation: Set5, Set14, BSD100, Urban100, and DIV2K-Val.

Table 8: Mean standard Y-channel performance over Set5, Set14, BSD100, Urban100, and DIV2K-Val, with scale-dependent border cropping. Each entry reports PSNR/SSIM. SwinIR-teacher is the teacher reference. **Bold** values indicate the best result among the distilled students.

Configuration	$2\times$ PSNR↑/SSIM↑	$3\times$ PSNR↑/SSIM↑	$4\times$ PSNR↑/SSIM↑
Baseline	33.3035/0.9215	29.4334/0.8307	27.2556/0.7599
SwinIR-teacher	35.1286/0.9387	30.5194/0.8656	28.4015/0.8065
KD	33.8241/0.9272	30.0279/0.8506	28.0896/0.7941
CSD	33.3181/0.9215	29.4684/0.8317	27.3416/0.7614
AugKD	33.2352/0.9210	29.3005/0.8287	27.2763/0.7638
DCKD	34.0042/0.9291	30.0928/0.8549	**28.2491** /0.8021
AMID	**34.1637/0.9306**	**30.1625/0.8574**	28.2480/ **0.8022**

At $2\times$, AMID improves over DCKD by 0.1595 dB in PSNR and 0.0015 in SSIM. At $3\times$, the corresponding gains are 0.0697 dB and 0.0025. At $4\times$, AMID and DCKD achieve nearly identical fidelity: DCKD is higher by only 0.0011 dB in PSNR, while AMID is higher by 0.0001 in SSIM. AMID also consistently outperforms KD, CSD, and AugKD at all three scales. These results confirm that the relative advantage of AMID remains consistent across RGB and standard Y-channel evaluation protocols.

A.7 Additional teacher–scale comparisons

Table 9 reports the mean performance across five standard benchmarks under multiple teacher architectures. KD and CSD bring only marginal gains over the baseline and occasionally perform worse. This confirms that output-level alignment alone cannot bridge the architectural gap in heterogeneous distillation. In contrast, AMID consistently achieves the best results on nearly all metrics across scales and teachers. At $2\times$, AMID delivers higher PSNR than both KD and CSD while producing noticeably better perceptual quality. For instance, SRFormerV2-ours reaches 67.79 MUSIQ with a PSNR that surpasses KD by 0.23 dB and CSD by 0.8 dB. The perceptual advantage grows at higher scales. At $3\times$, HAT-ours raises MUSIQ from 61.27 to 63.67 and reduces LPIPS from 0.2031 to 0.1925 compared with KD, whereas CSD variants stay close to the baseline. At $4\times$, MambaIRV2-ours lifts MUSIQ from 52.16 to 58.13 and cuts LPIPS from 0.3007 to 0.2769, nearly halving the perceptual gap to the teacher. Notably, even with the low-fidelity HYPIR teacher (25.68 dB PSNR at $2\times$, well below the 31.37 dB baseline), KD falls below the undistilled student, whereas AMID reaches 31.95 dB and remains stronger than KD and CSD on most metrics. The exceptions are CSD on $2\times$ MUSIQ and $4\times$ PSNR, CSD on $3\times$ LPIPS and CLIPIQA, and KD on $3\times$ MANIQA. Over-

all, these results show that AMID effectively decouples the student from the teacher architecture, reliably transferring both fidelity and perceptual priors in heterogeneous settings.

Table 9: Mean performance on five standard benchmarks. "baseline": student without distillation. "-ours": GAMD + ISCD. "-KD", "-CSD": KD (Hinton et al., 2015), CSD (Wang et al., 2021b) distillation. No suffix: teacher reference. ↑/↓: higher/lower better. **Bold** indicates the best among distilled students.

Scale	Configuration	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
2×	baseline	31.3700	0.9142	0.0996	31.0004	5.2242	0.6203	66.8814	0.4225
	MambaIRV2	33.6690	0.9348	0.0730	27.9410	4.9735	0.6690	67.8564	0.4606
	MambaIRV2-KD	31.8964	0.9206	0.0907	30.1636	5.1072	0.6385	67.2816	0.4344
	MambaIRV2-CSD	31.3508	0.9141	0.0963	31.0268	5.1743	0.6143	67.1616	0.4266
	MambaIRV2-ours	**32.1450**	**0.9232**	**0.0870**	**30.1181**	**5.0634**	**0.6402**	**67.3918**	**0.4410**
	SRFormerV2	33.5860	0.9350	0.0713	27.6579	4.9473	0.6623	67.9640	0.4637
	SRFormerV2-KD	31.9082	0.9204	0.0908	29.9472	5.0922	0.6352	67.3467	0.4372
	SRFormerV2-CSD	31.3369	0.9139	0.0965	31.2492	5.1970	0.6159	67.2578	0.4274
	SRFormerV2-ours	**32.1372**	**0.9235**	**0.0832**	**29.1614**	**4.9887**	**0.6392**	**67.7863**	**0.4474**
	HYPIR	25.6756	0.7840	0.1120	16.2458	4.5055	0.6812	67.8484	0.4627
	HYPIR-KD	30.8456	0.8986	0.1280	32.0777	5.3046	0.6019	66.6199	0.4195
	HYPIR-CSD	31.3647	0.9138	0.0963	31.0972	5.1608	0.6142	**67.3837**	0.4276
	HYPIR-ours	**31.9532**	**0.9211**	**0.0887**	**30.1585**	**5.0842**	**0.6322**	67.3193	**0.4374**
3×	baseline	27.6764	0.8170	0.2493	44.5242	6.7100	0.4994	54.8755	0.3314
	HAT	28.8218	0.8603	0.1650	36.7283	5.3469	0.5998	66.1524	0.4466
	HAT-KD	28.1231	0.8370	0.2031	41.2913	5.9057	0.5508	61.2670	0.3805
	HAT-CSD	27.7009	0.8185	0.2453	44.4854	6.9227	0.5009	55.3825	0.3342
	HAT-ours	**28.1907**	**0.8441**	**0.1925**	**39.8828**	**5.6760**	**0.5703**	**63.6696**	**0.4082**
	SRFormerV2	28.8723	0.8581	0.1689	37.4848	5.4006	0.5918	66.0950	0.4429
	SRFormerV2-KD	28.1358	0.8370	0.2041	41.2461	5.8923	0.5483	61.4001	0.3800
	SRFormerV2-CSD	27.6760	0.8181	0.2460	44.6336	6.8045	0.4995	55.2394	0.3335
	SRFormerV2-ours	**28.2180**	**0.8441**	**0.1917**	**39.9624**	**5.6303**	**0.5724**	**63.8683**	**0.4096**
	HYPIR	24.6417	0.7396	0.1638	21.7624	4.5913	0.6375	63.4968	0.4167
	HYPIR-KD	27.4746	0.8085	0.2642	45.4615	6.9325	0.4865	55.5948	**0.3371**
	HYPIR-CSD	27.6884	0.8179	**0.2442**	44.3883	6.6810	**0.5006**	55.5389	0.3343
	HYPIR-ours	**27.7458**	**0.8190**	0.2443	**43.6027**	**6.6770**	0.4995	**55.6383**	0.3365
4×	baseline	25.5859	0.7408	0.3497	51.3503	8.0077	0.4265	44.9011	0.2767
	HAT	26.4900	0.7935	0.2434	43.7787	5.7753	0.5478	63.4440	0.4295
	HAT-KD	26.1827	0.7739	0.2908	48.0659	6.6917	0.4785	53.3278	0.3296
	HAT-CSD	25.6518	0.7421	0.3522	51.8988	7.3310	0.4234	44.3845	0.2749
	HAT-ours	**26.2090**	**0.7764**	**0.2895**	**47.9402**	**6.5970**	**0.4790**	**54.0848**	**0.3338**
	MambaIRV2	27.0692	0.7993	0.2481	44.0909	5.9348	0.5386	62.4596	0.4167
	MambaIRV2-KD	26.2080	0.7745	0.3007	49.3713	6.6494	0.4889	52.1594	0.3239
	MambaIRV2-CSD	25.6388	0.7418	0.3529	51.5239	7.6018	0.4256	44.3387	0.2756
	MambaIRV2-ours	**26.2317**	**0.7832**	**0.2769**	**45.4262**	**6.0860**	**0.5184**	**58.1255**	**0.3675**
	HYPIR	23.7373	0.6958	0.2640	33.0728	5.9516	0.5324	49.6853	0.3370
	HYPIR-KD	25.6216	0.7422	0.3517	51.4405	7.6412	0.4233	45.6980	0.2851
	HYPIR-CSD	**25.6757**	0.7420	0.3556	51.2420	7.6050	0.4250	44.0919	0.2715
	HYPIR-ours	25.6596	**0.7456**	**0.3406**	**50.1306**	**7.5009**	**0.4351**	**48.9535**	**0.3118**

A.8 Transfer to an EDSR student architecture

The main experiments use a fixed RRDB student to isolate the effects of heterogeneous teachers and distillation objectives. We additionally replace RRDB with standard EDSR (Lim et al., 2017a) to examine whether AMID depends on a particular student architecture. The comparison in Table 10 uses a SwinIR teacher at $4\times$ and follows the same training and evaluation protocol as the main study.

ISCD requires the frozen inverse network to share the student backbone. For this experiment the inverse is therefore instantiated with the EDSR residual architecture, the input and output are reversed relative to the EDSR student, and the network is pretrained on the same HR-to-LR pairs under the inverse reconstruction objective in Appendix A.1.6. After pretraining, all inverse weights remain frozen during distillation. GAMD is left unchanged, since it depends only on model outputs and input attributions.

Standard EDSR is not a lightweight architecture and is not uniformly smaller than the SwinIR teacher. The experiment tests whether attribution matching and student-matched inverse alignment remain effective on a second CNN student. It does not evaluate large-to-small compression.

Table 10: Distillation results with a standard EDSR student and a SwinIR teacher at $4\times$. **Baseline**: EDSR trained without distillation. SwinIR denotes the teacher reference. ↑/↓: higher/lower better. **Bold** values indicate the best result among the distilled EDSR students.

Configuration	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
Baseline	25.2889	0.7319	0.3495	49.4686	7.1708	0.4542	44.5893	0.2699
SwinIR teacher	26.7267	0.7908	0.2528	43.9823	5.9344	0.5370	62.8920	0.4186
KD	**26.4244**	0.7741	0.2894	46.1007	6.7962	0.4998	55.1871	0.3315
CSD	26.3991	0.7729	0.2939	45.3911	6.7303	0.5052	54.4242	0.3281
AugKD	25.3169	0.7350	0.3424	49.3871	7.1303	0.4503	45.4553	0.2760
DCKD	26.2272	0.7727	0.2785	45.2307	6.2866	0.5286	58.2626	0.3487
AMID (ours)	26.3382	**0.7762**	**0.2765**	**44.2994**	**6.2355**	**0.5346**	**58.9899**	**0.3563**

AMID achieves the best performance among all distilled EDSR students on seven of eight metrics. Compared with the undistilled EDSR baseline, it improves PSNR by 1.05 dB, reduces LPIPS from 0.3495 to 0.2765, and raises MUSIQ from 44.59 to 58.99. KD achieves the highest PSNR among the distilled students, exceeding AMID by 0.09 dB, but AMID performs better on all other metrics.

AMID also outperforms DCKD on every reported metric, including improvements of 0.11 dB in PSNR, 0.73 in MUSIQ, and 0.0076 in MANIQA. The result indicates that the AMID losses remain effective when both the student and its inverse proxy are replaced by an EDSR pair. Deployment-efficiency conclusions in the main study continue to rest on the fixed RRDB student.

A.9 Complementary effects of GAMD and ISCD

The ablation in Table 11 isolates GAMD-only (w/G), ISCD-only (w/I), and the full combination. At $2\times$, GAMD alone raises PSNR by approximately 0.1 to 0.3 dB with only marginal movement on perceptual metrics, whereas ISCD alone produces substantially larger gains on every reported metric and is the primary source of both fidelity and perceptual improvement at this scale. At $3\times$ and $4\times$, ISCD again accounts for the dominant perceptual gains, while GAMD leaves PSNR nearly unchanged relative to the undistilled student and induces only small shifts in NIQE and MUSIQ. Under a SwinIR teacher at $4\times$, ISCD alone lifts MUSIQ from 44.90 to 59.62 and reduces LPIPS from 0.3497 to 0.2688, while GAMD alone changes these scores by less than 0.002. The combination attains the best overall result on most teacher-scale settings, with ISCD providing the principal perceptual supervision and GAMD acting mainly as a fidelity regularizer. Two settings depart from this pattern. With HAT at $4\times$, the ISCD-only student is stronger than the combination on LPIPS, BRISQUE, NIQE, CLIPIQA, MUSIQ, and MANIQA, while the combination is higher only on PSNR and SSIM. With HYPIR at $3\times$, the GAMD-only student attains better scores on all reported perceptual metrics, including MUSIQ of 56.62 against 55.64 for the combination. The combination is higher on PSNR and is tied with GAMD-only on SSIM. These cases indicate that introducing the second term can improve fidelity while lowering selected perceptual scores, and that the combination need not raise every metric at once. Under HYPIR at $2\times$ and $4\times$, the combination remains stronger than either component alone, and GAMD-only still exceeds the undistilled $2\times$ baseline (31.62 versus 31.37 dB) despite the low teacher PSNR.

Fig. 7 shows a representative crop that visualizes the same component separation (baseline, w/G, w/I, and full AMID).

Table 11: Mean performance on five standard benchmark datasets. “baseline” indicates the baseline student model trained without any distillation from a teacher model. For entries in the “Configuration” column, a suffix of “ours” indicates that the student is distilled via our pipeline(GAMD + ISCD) using the specified teacher architecture. A suffix of “w/G” denotes distillation only via GAMD and “w/I” denotes distillation only via ISCD. ↑/↓: higher/lower better. **Bold** indicates the best performance among all configurations.

Scale	Configuration	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
2×	baseline	31.3700	0.9142	0.0996	31.0004	5.2242	0.6203	66.8814	0.4225
	SwinIR-w/G	31.5413	0.9165	0.0945	30.8690	5.1605	0.6248	67.1393	0.4272
	SwinIR-w/I	31.9716	0.9214	0.0866	**29.5424**	5.0599	0.6327	67.6019	0.4412
	SwinIR-ours	**32.2072**	**0.9240**	**0.0860**	29.6972	**5.0277**	**0.6375**	**67.6334**	**0.4450**
	MambaIRV2-w/G	31.4729	0.9151	0.0981	30.9629	5.2379	0.6234	66.9100	0.4226
	MambaIRV2-w/I	31.9557	0.9210	0.0886	**30.1076**	5.0813	0.6384	67.2937	0.4369
	MambaIRV2-ours	**32.1450**	**0.9232**	**0.0870**	30.1181	**5.0634**	**0.6402**	**67.3918**	**0.4410**
	SRFormerV2-w/G	31.5321	0.9163	0.0959	30.6500	5.2149	0.6286	66.9718	0.4246
	SRFormerV2-w/I	31.9400	0.9210	0.0874	29.6413	5.0642	0.6345	67.4953	0.4402
	SRFormerV2-ours	**32.1372**	**0.9235**	**0.0832**	**29.1614**	**4.9887**	**0.6392**	**67.7863**	**0.4474**
	HYPIR-w/G	31.6216	0.9170	0.0940	30.5994	5.1271	0.6243	67.1056	0.4296
	HYPIR-w/I	31.7675	0.9183	0.0928	30.3341	**5.0543**	0.6264	67.3133	0.4310
	HYPIR-ours	**31.9532**	**0.9211**	**0.0887**	**30.1585**	5.0842	**0.6322**	**67.3193**	**0.4374**
3×	baseline	27.6764	0.8170	0.2493	44.5242	6.7100	0.4994	54.8755	0.3314
	SwinIR-w/G	27.7027	0.8182	0.2449	44.2413	6.6338	0.5038	55.6986	0.3350
	SwinIR-w/I	28.1699	0.8424	0.1926	40.2894	5.6425	0.5672	63.5977	0.4095
	SwinIR-ours	**28.2295**	**0.8453**	**0.1875**	**39.4604**	**5.5506**	**0.5712**	**64.5399**	**0.4206**
	HAT-w/G	27.7101	0.8183	0.2455	44.4977	6.7893	0.5013	55.2691	0.3330
	HAT-w/I	28.1285	0.8407	0.1976	40.7580	5.7502	0.5648	62.6522	0.3958
	HAT-ours	**28.1907**	**0.8441**	**0.1925**	**39.8828**	**5.6760**	**0.5703**	**63.6696**	**0.4082**
	SRFormerV2-w/G	27.7109	0.8188	0.2417	43.6096	6.5968	0.5044	56.1812	0.3380
	SRFormerV2-w/I	28.1363	0.8411	0.1954	40.5316	5.6650	0.5690	63.1487	0.4014
	SRFormerV2-ours	**28.2180**	**0.8441**	**0.1917**	**39.9624**	**5.6303**	**0.5724**	**63.8683**	**0.4096**
	HYPIR-w/G	27.6925	**0.8190**	**0.2392**	**43.4310**	**6.5554**	**0.5057**	**56.6236**	**0.3438**
	HYPIR-w/I	27.6876	0.8176	0.2449	43.9110	6.6357	0.5002	55.7559	0.3360
	HYPIR-ours	**27.7458**	**0.8190**	0.2443	43.6027	6.6770	0.4995	55.6383	0.3365
4×	baseline	25.5859	0.7408	0.3497	51.3503	8.0077	0.4265	44.9011	0.2767
	SwinIR-w/G	25.6718	0.7427	0.3483	51.3903	7.5867	0.4273	45.1113	0.2766
	SwinIR-w/I	26.2567	0.7833	0.2688	45.0991	5.9216	0.5143	59.6175	0.3815
	SwinIR-ours	**26.3335**	**0.7855**	**0.2669**	**45.0321**	**5.9207**	**0.5174**	**60.2025**	**0.3870**
	HAT-w/G	25.6004	0.7421	0.3461	51.1292	7.5388	0.4267	45.5459	0.2817
	HAT-w/I	26.1969	0.7754	**0.2881**	**47.9308**	**6.5597**	**0.4840**	**54.4313**	**0.3354**
	HAT-ours	**26.2090**	**0.7764**	0.2895	47.9402	6.5970	0.4790	54.0848	0.3338
	MambaIRV2-w/G	25.5926	0.7410	0.3490	51.2525	7.4188	0.4279	44.8683	0.2765
	MambaIRV2-w/I	**26.2353**	0.7821	0.2779	45.7632	6.0993	0.5126	57.7399	0.3594
	MambaIRV2-ours	26.2317	**0.7832**	**0.2769**	**45.4262**	**6.0860**	**0.5184**	**58.1255**	**0.3675**
	HYPIR-w/G	25.6246	0.7415	0.3498	51.0994	**7.4735**	0.4298	44.8073	0.2769
	HYPIR-w/I	25.6264	0.7447	0.3455	**50.0120**	7.5800	0.4266	48.3149	0.3034
	HYPIR-ours	**25.6596**	**0.7456**	**0.3406**	50.1306	7.5009	**0.4351**	**48.9535**	**0.3118**

A.10 Ablation study of GAMD

This appendix discusses the attribution guidance map $\mathbf{U}$ in Global Attribution Map Distillation (GAMD). We explain how the weight $u(\hat{R})$ depends on batch-normalized residuals $\hat{R}$, why the map is U-shaped, and how much each guidance policy contributes when the rest of GAMD is unchanged.

The residual $\mathbf{R}$, the normalized residual $\hat{R}$, and the weight function $u(\hat{R})$ are defined in Sec. 3.1 (*Attribution guidance map estimation*) and Appendix A.1.3. The guidance-weighted response and the distillation loss appear in equation 3 and equation 5, respectively. Fig. 3 plots the three policies used in ablation (Uniform, Monotonic, and Ours). Fig. 5 shows a spatial example of $\mathbf{R}$, $\hat{R}$, and $\mathbf{U}$.

The subsequent subsections develop the rationale for tail-selective guidance in heterogeneous super-resolution and report map ablations in which the mapping $\hat{R} \mapsto u(\hat{R})$ is the only modified component. In these experiments, the student is optimized with $\mathcal{L}_{\mathrm{REC}} + \mathcal{L}_{\mathrm{GAMD}}$(-w/G) while ISCD is disabled, supervision is provided by a frozen teacher, and performance is averaged over the five RGB benchmarks at each scale defined in Table 3 and Table 11. Table 14 supplements the Uniform, Monotonic, and Ours guidance maps with the $\mathcal{L}_{\mathrm{REC}}$-only baseline, thereby separating the effect of introducing GAMD from that of the inverted-Gaussian weighting.

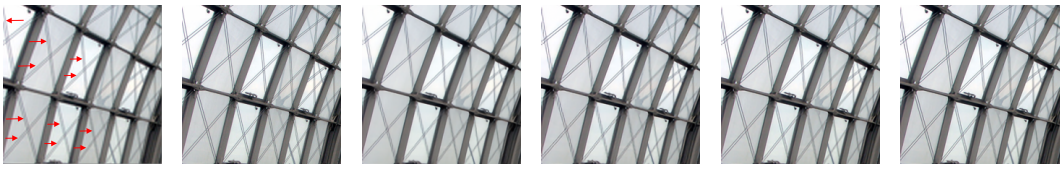


Figure 7: Component ablation on a representative $2\times$ patch (teacher: HAT). **LR**: low-resolution input; **GT**: ground-truth high-resolution reference (not a model output).**Baseline**: undistilled student; **w/G**: GAMD only; **w/I**: ISCD only; **ours**: GAMD + ISCD. Differences in fine details are best viewed when the figure is zoomed in.

A.10.1 Guidance map as attribution weights

Each element of the guidance map, $U_{n,1,h,w} = u(\hat{R}_{n,1,h,w}) \in [0,1]$, specifies an *attribution guidance weight* at the corresponding HR location. The weight scales that location in the guidance-weighted structural response equation 3 and, through path-integrated gradients, in the GAMD objective equation 5. Locations with large u are those at which matching $\Phi^S(x)$ to $\Phi^T(x)$ is expected to change how the student uses LR evidence. Locations with small u add little supervision and can divert gradient mass away from batch positions where the teacher-student attribution gap is more informative.

Accordingly, $\mathbf{U}$ selects spatial sites inside the current mini-batch at which aligning the student's input-utilization behavior with the teacher is most useful. It is not a measure of predictive uncertainty, per-pixel confidence, or the probability of reconstruction error. Ground-truth fidelity is enforced by $\mathcal{L}_{\mathrm{REC}}$. Local reconstruction error enters $\mathbf{U}$ only through $\mathbf{R}$ and its batch-normalized form $\hat{R}$.

Integrated-gradient attributions in LAM (Gu & Dong, 2020) describe how each input contributes to each output, and they do not measure agreement with the ground truth. Recent work on attribution and model editing likewise treats reasoning strategy as the transferable quantity (Chen et al., 2024). In GAMD, $\mathbf{U}$ weights the output-gradient field that defines $\mathcal{R}(x)$ and thereby selects the sites that enter the matched attribution target equation 2-equation 5. The map is computed from the student prediction and the ground truth during training and is unused at inference.

A.10.2 Attribution supervision under small residuals

In heterogeneous distillation, a student can match the ground truth closely while following a different internal reasoning path from the teacher. This is the familiar *right answer, wrong reason* setting: agreement in HR appearance does not guarantee agreement in how LR evidence is used. Our student is RRDB with local CNN pathways, whereas teachers span Transformers (SwinIR, HAT) (Liang et al., 2021; Chen et al., 2025), state-space models (MambaIRV2) (Guo et al., 2025), hybrid designs (SRFormerV2) (Zhou et al., 2024b), and diffusion-style restorers (HYPIR) (Lin et al., 2025).

Table 12 summarizes typical pixel categories after batch-wise normalization of $\mathbf{R}$ to $\hat{R}$. Appendix A.10.4 then specifies the U-shaped guidance policy that assigns supervision strength to each category. Output-only matching is fragile under distribution shift when the reasoning process is left unmatched (Zagoruyko & Komodakis, 2016; Taneja & Banerjee, 2026). AMID targets on-device deployment, so the student is trained to inherit the teacher's input-utilization behavior and is not trained to copy teacher pixels alone.

Small residual and guidance weight Batch min-max normalization maps $\mathbf{R}$ to $\hat{R} \in [0,1]$ within each batch. The tail $\hat{R} \to 0$ therefore marks pixels that are *best reconstructed relative to other pixels in the same batch*, and it does not mark pixels whose routing policy already matches the teacher. Super-resolution is ill-posed: many HR images are consistent with a single LR input, so teacher and student may adopt different LR routing policies while still yielding similar HR appearance.

Table 12: Informal pixel categories under batch-normalized $\hat{R}$ and intended GAMD supervision under the **Ours** guidance map. Categories are defined by residual rank inside the mini-batch, not by semantic region type.

Category	Student vs. GT	Teacher vs. GT	Student reasoning	Teacher reasoning	GAMD action
Low-residual (easy tail)	Closest in-batch	Typically close	May rely on local texture matching (CNN bias)	May use long-range context (attention or state-space mixing)	Align attribution. Appearance can agree while the LR routing path still differs.
Mid-residual (homogeneous)	Near the batch median	Typically comparable	Shared interpolation-like upsampling	Shared simple strategy	Mute. Little additional reasoning gap.
High-residual (failed tail)	Farthest in-batch	Often closer than the student	Wrong or weak routing	Teacher-preferred routing	Align attribution. This tail is the primary teaching signal.

Cross-architecture example. A Transformer teacher may recover texture from distant self-similar LR patches, whereas RRDB may rely on nearby dense connections. Both outputs can appear correct relative to the ground truth while $\Phi^S \not\approx \Phi^T$. Attribution alignment therefore supervises the teacher's decision-making strategy. This principle aligns with hint and attention transfer (Romero et al., 2014; Zagoruyko & Komodakis, 2016), but GAMD applies it to attribution maps Φ obtained from integrated gradients of $\mathcal{R}(x)$, without requiring aligned hidden layers between heterogeneous architectures. The two batch tails call for high guidance for complementary reasons. At $\hat{R} \to 0$, the HR result may be good yet the reasoning path may still diverge from the teacher. At $\hat{R} \to 1$, the HR result is poor and directly indicates routing failure. Monotonic maps that increase the guidance weight only with $\hat{R}$ suppress the low tail and cannot correct strategy mismatch hidden behind good ground-truth fit (Table 12).

A.10.3 Down-weighting of mid-range residuals

Mid-range residuals ($\hat{R} \approx 0.5$) denote pixels whose reconstruction error sits near the batch median. They often occupy large spatially homogeneous regions in which the residual varies slowly, including weakly mismatched smooth interiors and weakly structured areas. These locations form a high-mass mid-range bulk. Well-reconstructed flat regions belong to the easy tail $\hat{R} \to 0$ and are not part of this bulk.

In the mid-range bulk, the student and the teacher often apply similar upsampling-like strategies, so Φ^S and Φ^T are already largely aligned. Additional $\mathcal{L}_{\mathrm{GAMD}}$ therefore contributes little marginal benefit, because there is no meaningful reasoning gap to close. If this homogeneous mass is overweighted, gradients for the two informative tails are diluted, since $\mathcal{R}(x)$ aggregates weighted output-gradient contributions and the mid-range bulk then dominates the attribution loss.

The same down-weighting principle appears in focal loss for easy classification examples (Lin et al., 2017), and here it is applied in attribution space to the mid-range residual mass. Sec. 3.1 motivates a non-monotonic guidance map that de-emphasizes this mid-range bulk and emphasizes both batch-relative extremes. Appendix A.10.4 specifies that map.

A.10.4 Inverted-Gaussian tail-selective guidance

After batch-wise normalization, $\hat{R} \in [0, 1]$ ranks pixels by **relative** reconstruction difficulty within the mini-batch. We map $\hat{R}$ to guidance weights with an inverted Gaussian ($\mu = 0.5$, $\sigma = 1/6$, Appendix A.1.3) and use the resulting map to modulate the attribution target in equation 2-equation 5.

- **Low** u at $\hat{R} \approx 0.5$. Homogeneous moderate residuals near the batch median.

- **High** u at $\hat{R} \to 0$ and $\hat{R} \to 1$. The two in-batch extremes, where teacher-student attribution gaps remain informative (Table 12).

Monotonic guidance maps over-weight high-error mass and suppress the low tail (Appendix A.10.2). Fig. 3 contrasts Uniform ($u \equiv 1$), Monotonic, and Ours policies on $\hat{R} \in [0, 1]$. Table 13 summarizes the rank-level supervision rules and the expected effect of each policy.

Table 13: Rank-level supervision induced by alternative guidance maps under a shared $\hat{R}$ pipeline. Rows are residual ranks, not semantic region types.

Residual rank	Uniform	Monotonic	Ours
Low-residual easy tail ($\hat{R} \to 0$)	✓	×	✓
Mid-residual homogeneous bulk ($\hat{R} \approx 0.5$)	✓	✓	×
High-residual failed tail ($\hat{R} \to 1$)	✓	✓	✓
Selection rule	None (equal weight)	Larger $\hat{R} \Rightarrow$ higher u	Mid $\hat{R}$ low, both extremes high
Expected effect	Diluted focus on informative tails	Over-weights high-error mass and drops the easy tail	Selective focus on both informative extremes

The inverted Gaussian is a **saliency filter**, not a probabilistic confidence model: smooth, low at mid $\hat{R}$, high at both tails, differentiable for integrated-gradient (IG) attribution training. Other smooth U-shaped filters (parabolic, raised cosine) implement the same policy class; we fix the Gaussian form for reproducibility. With $u(\hat{R}) = 1 - \bar{g}(\hat{R})$, $u(0) = u(1) = 1$, $u(0.5) = 0$, and $u(\hat{R}) = u(1 - \hat{R})$. Symmetry in $\hat{R}$ is deliberate rank-based tail selection, not a claim that both tails are semantically identical.

A.10.5 Guidance-map ablation

Protocol. RRDB student, frozen HAT teacher, $\mathcal{L}_{\mathrm{REC}} + \mathcal{L}_{\mathrm{GAMD}}$ (ISCD off). Training: DF2K, 384×384 HR crops, batch size 8, Adam 10^{-4}, 250,000 iterations, scales $2\times$ /$3\times$ /$4\times$. Path-integrated gradients use the same step count and GAMD frequency as full AMID. Evaluation: mean over Set5, Set14, BSD100, Urban100, and DIV2K-Val (RGB space). Only the guidance map $u(\cdot)$ varies across ablated rows.

Results. Table 14 reports five-benchmark means for a frozen HAT teacher under the GAMD-only protocol. The **Baseline** entry uses $\mathcal{L}_{\mathrm{REC}}$ alone and reproduces the undistilled student in Table 3. **GAMD+Uniform**, **GAMD+Monotonic**, and **GAMD+Ours** share the same HAT teacher, residual $\mathbf{R}$, normalized residual $\hat{R}$, integrated-gradient schedule, and $\mathcal{L}_{\mathrm{GAMD}}$, and they differ only in the guidance map $u(\cdot)$. **GAMD+Ours** matches HAT-w/G in Table 11 at all three scales and matches the $2\times$ HAT-w/G row in Table 3. All w/G entries use the inverted-Gaussian map.

Across $2\times$, $3\times$, and $4\times$, **GAMD+Ours** attains the highest PSNR, SSIM, LPIPS, MUSIQ, and MANIQA among the compared configurations. It also leads on BRISQUE, NIQE, and CLIPIQA at $2\times$. At $3\times$ and $4\times$, **GAMD+Monotonic** obtains slightly better scores on isolated no-reference metrics, including BRISQUE at both scales, NIQE at $3\times$, and CLIPIQA at $4\times$. The margins remain small, and these no-reference scores can diverge from full-reference fidelity. PSNR, SSIM, LPIPS, MUSIQ, and MANIQA are therefore used as the primary evidence for map design, and the remaining no-reference splits are treated as secondary.

Two comparisons isolate the contribution of GAMD and of the tail-selective map. Relative to **Baseline**, **GAMD+Ours** improves PSNR, SSIM, LPIPS, MUSIQ, and MANIQA at every scale, which indicates that attribution matching transfers teacher signal beyond reconstruction supervision. Relative to **GAMD+Uniform**, **GAMD+Ours** raises PSNR from 31.59 to 31.64 at $2\times$, from 27.61 to 27.71 at $3\times$, and from 25.56 to 25.60 at $4\times$, with consistent gains on the same primary metrics. The

inverted-Gaussian map is therefore an integral component of GAMD and is not an optional refinement of unweighted attribution matching. GAMD-only w/G results for the remaining teachers are reported in Table 3 and Table 11.

Table 14: Attribution guidance map ablation (HAT teacher when GAMD is enabled; five-benchmark mean). **Baseline**: $\mathcal{L}_{\mathrm{REC}}$ only (Table 3). **GAMD+Ours**: same as HAT-w/G in Table 3 and Table 11. **GAMD+Uni**: GAMD with uniform guidance map. **GAMD+Mono**: GAMD with monotonic guidance map. ↑/↓: higher/lower better. **Bold**: best among GAMD-enabled guidance maps at each scale.

Scale	Configuration	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
2×	Baseline	31.3700	0.9142	0.0996	31.0004	5.2242	0.6203	66.8814	0.4225
	GAMD+Uni	31.5912	0.9162	0.0978	31.0930	5.1639	0.6217	66.6700	0.4227
	GAMD+Mono	31.6227	0.9170	0.0953	30.4818	5.1516	0.6238	66.7599	0.4257
	GAMD+Ours	**31.6368**	**0.9175**	**0.0944**	**30.2326**	**5.1366**	**0.6295**	**66.9866**	**0.4277**
3×	Baseline	27.6764	0.8170	0.2493	44.5242	6.7100	0.4994	54.8755	0.3314
	GAMD+Uni	27.6067	0.8150	0.2520	44.3838	6.8273	0.4929	54.5817	0.3282
	GAMD+Mono	27.6647	0.8168	0.2483	**44.3250**	**6.7880**	0.4999	55.1869	0.3319
	GAMD+Ours	**27.7101**	**0.8183**	**0.2455**	44.4977	6.7893	**0.5013**	**55.2691**	**0.3330**
4×	Baseline	25.5859	0.7408	0.3497	51.3503	8.0077	0.4265	44.9011	0.2767
	GAMD+Uni	25.5634	0.7404	0.3559	50.9912	7.6609	0.4273	44.3241	0.2743
	GAMD+Mono	25.4648	0.7388	0.3503	**50.4110**	7.5638	**0.4309**	45.0309	0.2814
	GAMD+Ours	**25.6004**	**0.7421**	**0.3461**	51.1292	**7.5388**	0.4267	**45.5459**	**0.2817**

A.11 Attribution-correspondence controls for GAMD

The guidance-map ablation examines where attribution alignment should be emphasized. We further investigate whether GAMD depends on the correspondence between each student sample and the spatial structure of its teacher attribution target. All experiments (in Table 15) use a frozen HAT teacher, the inverted-Gaussian guidance map, and the same GAMD-only training protocol as Table 14.

Cross-sample applies a derangement within each mini-batch, matching the student attribution of sample i with the teacher attribution of a different sample $j \neq i$. This operation preserves the distribution of teacher attribution values while removing sample-level correspondence. **Spatial** randomly permutes positions within each teacher attribution map, preserving its value distribution while destroying its spatial organization. In both controls, the student attribution, guidance map, loss weight, and training schedule remain unchanged.

Table 15: GAMD attribution-correspondence controls with a HAT teacher. **Baseline**: reconstruction training without GAMD. **Matched**: GAMD with the correct teacher attribution target and inverted-Gaussian guidance. **Cross-sample**: teacher attribution targets are permuted across samples. **Spatial**: positions are permuted within each teacher attribution map. ↑/↓: higher/lower better.

Scale	Configuration	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
2×	Baseline	31.3700	0.9142	0.0996	31.0004	5.2242	0.6203	66.8814	0.4225
	Matched	31.6368	0.9175	0.0944	30.2326	5.1366	0.6295	66.9866	0.4277
	Cross-sample	31.3197	0.9142	0.0972	31.1779	5.2564	0.6179	67.3252	0.4280
	Spatial	31.3582	0.9139	0.0967	30.8384	5.1859	0.6168	67.1843	0.4258
3×	Baseline	27.6764	0.8170	0.2493	44.5242	6.7100	0.4994	54.8755	0.3314
	Matched	27.7101	0.8183	0.2455	44.4977	6.7893	0.5013	55.2691	0.3330
	Cross-sample	27.6043	0.8154	0.2536	44.5080	6.7303	0.4984	54.2749	0.3278
	Spatial	27.6374	0.8166	0.2520	44.5669	6.7139	0.4990	54.5211	0.3296
4×	Baseline	25.5859	0.7408	0.3497	51.3503	8.0077	0.4265	44.9011	0.2767
	Matched	25.6004	0.7421	0.3461	51.1292	7.5388	0.4267	45.5459	0.2817
	Cross-sample	25.5713	0.7399	0.3509	51.4856	8.2618	0.4263	44.5518	0.2746
	Spatial	25.5398	0.7400	0.3469	50.9203	8.2098	0.4277	45.3814	0.2780

Matched attribution consistently provides the strongest reconstruction fidelity. At 2×, it improves PSNR by 0.27 dB over the baseline, whereas both corrupted controls fall below the baseline. At 3×, Cross-sample and Spatial reduce PSNR to 27.6043 and 27.6374 dB, respectively, and also degrade

SSIM and LPIPS relative to the baseline. In contrast, Matched improves PSNR to 27.7101 dB and outperforms both shuffled controls on PSNR, SSIM, LPIPS, BRISQUE, CLIPIQA, MUSIQ, and MANIQA. NIQE is the only exception at this scale. At $4\times$, Matched again achieves the highest PSNR and SSIM and is the only GAMD configuration that jointly improves both metrics over the baseline.

Across the three scales, Matched outperforms Cross-sample in 21 of 24 metric–scale comparisons and Spatial in 20 of 24 comparisons. The remaining exceptions are limited to isolated perceptual or no-reference metrics. Cross-sample slightly favors MUSIQ and MANIQA at $2\times$, while Spatial favors MUSIQ at $2\times$ and BRISQUE and CLIPIQA at $4\times$. Crucially, disrupting either sample-level correspondence or spatial organization consistently removes the PSNR and SSIM gains of GAMD. These results demonstrate that GAMD benefits from the correct sample-specific and spatially structured teacher attribution target rather than from attribution regularization alone.

A.12 Attribution maps at different scales

We further investigated the model's local attribution patterns on structured and textured scenes. As shown in Fig 8, Fig 9 and Fig 10, the computed attribution maps reveal that GAMD expands the effective receptive field, enabling each query to access substantially more distant pixels than the baseline. This broader scope does not result in indiscriminate aggregation. When the query is centered on a region dominated by horizontal repetitive edge textures, GAMD preferentially draws information from distant pixels that share similar horizontal patterns. It does not blindly include all pixels within the enlarged window. This behavior indicates an implicit selection mechanism that purposely recruits beneficial contextual cues while filtering out irrelevant or harmful signals. Such targeted aggregation aligns with the method's core design and contributes to the improved reconstruction quality observed in our experiments. In summary, GAMD is an efficient and robust knowledge distillation strategy which can comprehensively and consistently elevate a compact student model's fidelity and perceptual quality across diverse teacher architectures and quality levels.

A.13 Ablation study of ISCD loss components

This appendix disentangles the three terms in the ISCD objective and complements the ISCD-only rows in Table 3 and Table 11. The main text presents ISCD as layer-wise contrastive alignment in a frozen student-isomorphic inverse network $\mathcal{I}$ (Sec. 3.2), together with output distillation and a degradation-cycle regularizer. A central empirical question is whether ISCD gains arise primarily from the proposed contrastive alignment or from conventional output matching and cycle regularization. A related question is whether improvements reflect additional deploy-time capacity rather than the intended supervision mechanism. For low-fidelity teachers such as HYPIR, it is also necessary to determine whether gains above the RRDB baseline are truly teacher dependent or can be explained by ground-truth supervision alone. We address these questions through a progressive loss ablation under a fixed ISCD-only protocol.

All runs use the same RRDB student, DF2K training schedule, five-benchmark evaluation, and $\mathcal{L}_{\mathrm{REC}}$, without GAMD. The distilled model remains the same 16.7 M RRDB in every configuration, and frozen $\mathcal{I}$ adds no trainable parameters to the deployed student. We therefore isolate the internal design of ISCD rather than the full AMID pipeline.

ISCD contains three complementary terms: output distillation $\mathcal{L}_{\mathrm{out}}$, inverse reconstruction $\mathcal{L}_{\mathrm{irc}}$, and multi-layer contrastive alignment $\mathcal{L}_{\mathrm{isc}}$. Their raw magnitudes differ substantially, so λ_{out}, λ_{irc}, and λ_{isc} are chosen once per teacher to keep the three weighted contributions comparable, following Table 6. The same values are used at $2\times$, $3\times$, and $4\times$ and remain fixed throughout training.

Progressive ablations of a given teacher reuse these coefficients. Disabling a component removes only that term, and the remaining coefficients stay unchanged. This protocol isolates each loss without retuning the retained components.

The output term $\mathcal{L}_{\mathrm{out}}$ is KD-style distillation on HR outputs and matches standalone KD in our implementation. The cycle term $\mathcal{L}_{\mathrm{irc}} = \|\mathcal{I}(\mathcal{S}(x)) - x\|_1$ imposes the degradation-cycle prior through fixed $\mathcal{I}$. The contrastive term $\mathcal{L}_{\mathrm{isc}}$ aligns $\mathcal{I}(\mathcal{S}(x))$ and $\mathcal{I}(\mathcal{T}(x))$ at corresponding layers and constitutes the proposed core transfer mechanism. ISCD therefore performs feature supervision in a

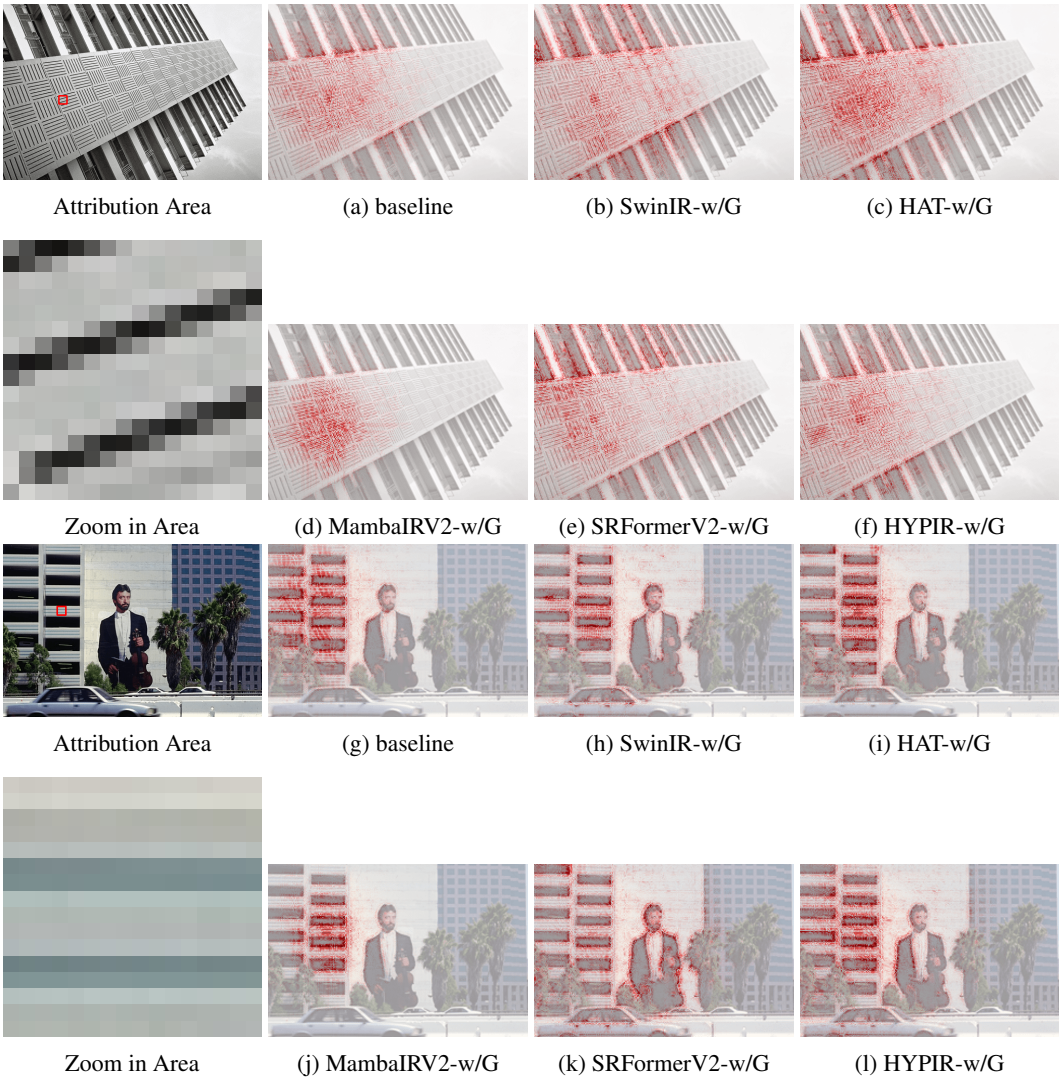


Figure 8: Local attribution maps at $2\times$.

student-native inverse proxy space; the present study tests the contribution of each term within that design.

For each teacher and scale, we accumulate losses in four steps: (1) $\mathcal{L}_{\text{out}}$ only; (2) $\mathcal{L}_{\text{out}} + \mathcal{L}_{\text{irc}}$; (3) $\mathcal{L}_{\text{out}} + \mathcal{L}_{\text{isc}}$; (4) full ISCD with all three terms. Tables 16 and 17 report five-benchmark means for HAT and HYPIR at $2\times$ and $4\times$. HAT serves as a high-fidelity Transformer teacher. HYPIR serves as a low-fidelity diffusion teacher with strong perceptual priors but only 25.68 dB PSNR at $2\times$ in Table 9.

For HYPIR at $2\times$, $\mathcal{L}_{\text{out}}$ alone yields 30.85 dB PSNR, which is below the 31.37 dB RRDB baseline. This indicates that naive output matching from this teacher cannot account for the final student performance by itself. Adding $\mathcal{L}_{\text{irc}}$ improves PSNR to 31.09 dB and reduces LPIPS from 0.1280 to 0.1011, but still does not exceed the baseline. Substantial gains appear only after introducing $\mathcal{L}_{\text{isc}}$, which requires teacher outputs through $\mathcal{I}(\mathcal{T}(x))$. The configuration $\mathcal{L}_{\text{out}} + \mathcal{L}_{\text{isc}}$ reaches 31.72 dB PSNR, 0.0954 LPIPS, 67.26 MUSIQ, and 0.4310 MANIQA. Relative to $\mathcal{L}_{\text{out}}$ alone, this corresponds to approximately $+0.88$ dB PSNR, -0.033 LPIPS, $+0.64$ MUSIQ, and $+0.012$ MANIQA. Full ISCD further improves to 31.77 dB PSNR and 0.0928 LPIPS. Teacher-dependent transfer above baseline therefore requires layer-aligned contrastive supervision rather than ground-truth reconstruction or cycle regularization alone.

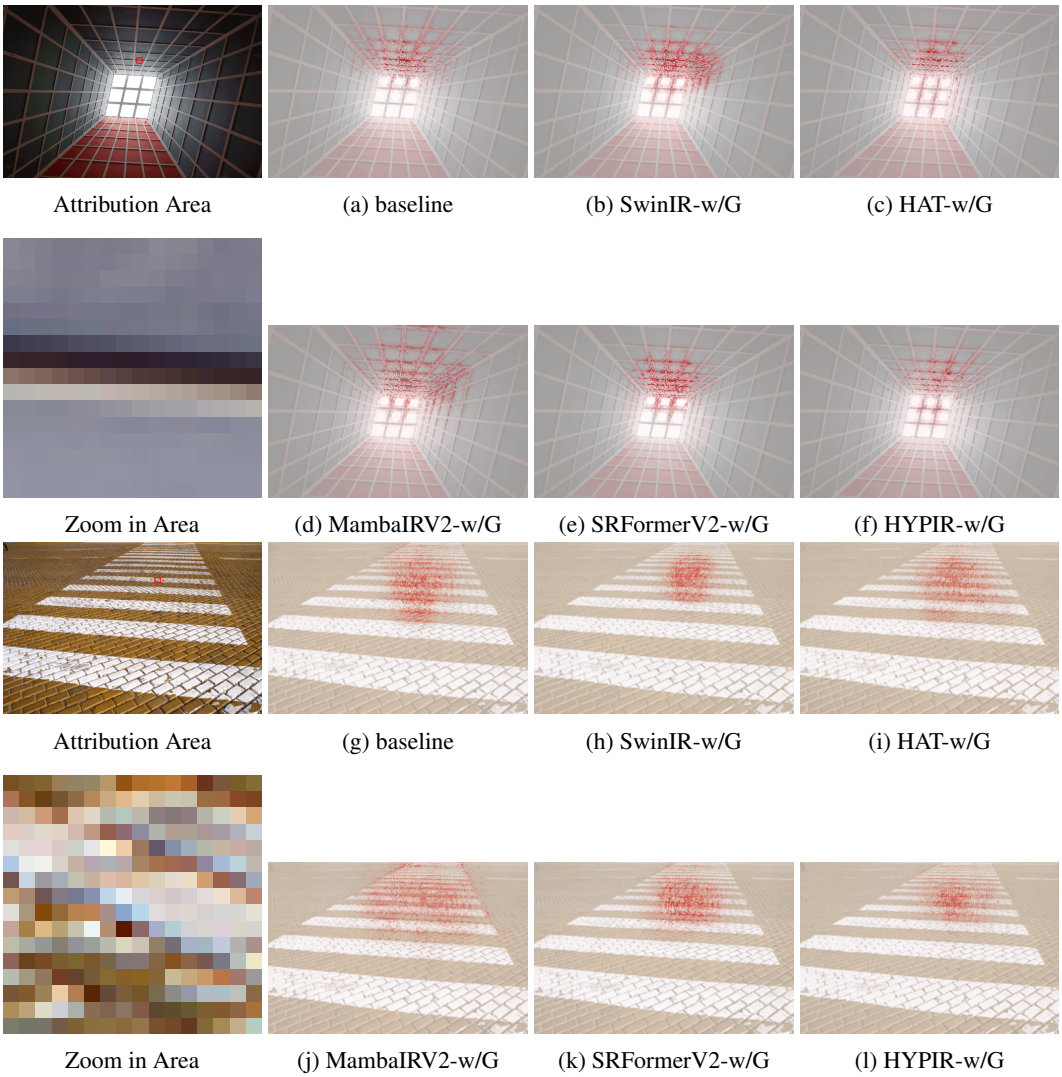


Figure 9: Local attribution maps at $3\times$.

For HAT at $2\times$, $\mathcal{L}_{\text{out}}$ gives 31.89 dB PSNR, consistent with HAT-KD in Table 2. Adding $\mathcal{L}_{\text{isc}}$ raises PSNR to 31.95 dB and lowers LPIPS to 0.0901. Full ISCD reaches 31.97 dB PSNR and 0.0877 LPIPS. The progression is monotonic or near-monotonic across metrics, indicating that layer-wise inverse alignment provides incremental benefit even when the teacher already supports a strong KD baseline.

At $4\times$, PSNR changes are smaller, but perceptual gains from $\mathcal{L}_{\text{isc}}$ remain clear. For HAT, $\mathcal{L}_{\text{out}}$ attains 26.18 dB PSNR and 53.33 MUSIQ. Adding $\mathcal{L}_{\text{irc}}$ raises MUSIQ to 53.96 with negligible PSNR change. Adding $\mathcal{L}_{\text{isc}}$ increases MUSIQ to 54.29 and MANIQA to 0.3344, which is approximately $+0.96$ MUSIQ and $+0.005$ MANIQA over $\mathcal{L}_{\text{out}}$ alone. Full ISCD reaches 26.20 dB PSNR and 54.43 MUSIQ. For HYPIR at $4\times$, the configuration $\mathcal{L}_{\text{out}} + \mathcal{L}_{\text{isc}}$ improves MUSIQ from 45.70 to 48.34 and MANIQA from 0.285 to 0.304 over output distillation alone. Full ISCD yields the best overall balance at 25.63 dB PSNR, 0.3455 LPIPS, and 48.31 MUSIQ, and improves over $\mathcal{L}_{\text{out}}$ on all eight reported metrics.

Three conclusions follow from the progressive ablation. First, $\mathcal{L}_{\text{out}}$ defines a KD baseline within ISCD but is not sufficient for strong heterogeneous transfer on its own, especially for HYPIR at $2\times$. Second, $\mathcal{L}_{\text{isc}}$ is the primary driver of teacher-dependent gains, because consistent improvements above baseline or above KD appear only when contrastive alignment through $\mathcal{I}$ is enabled. Third,

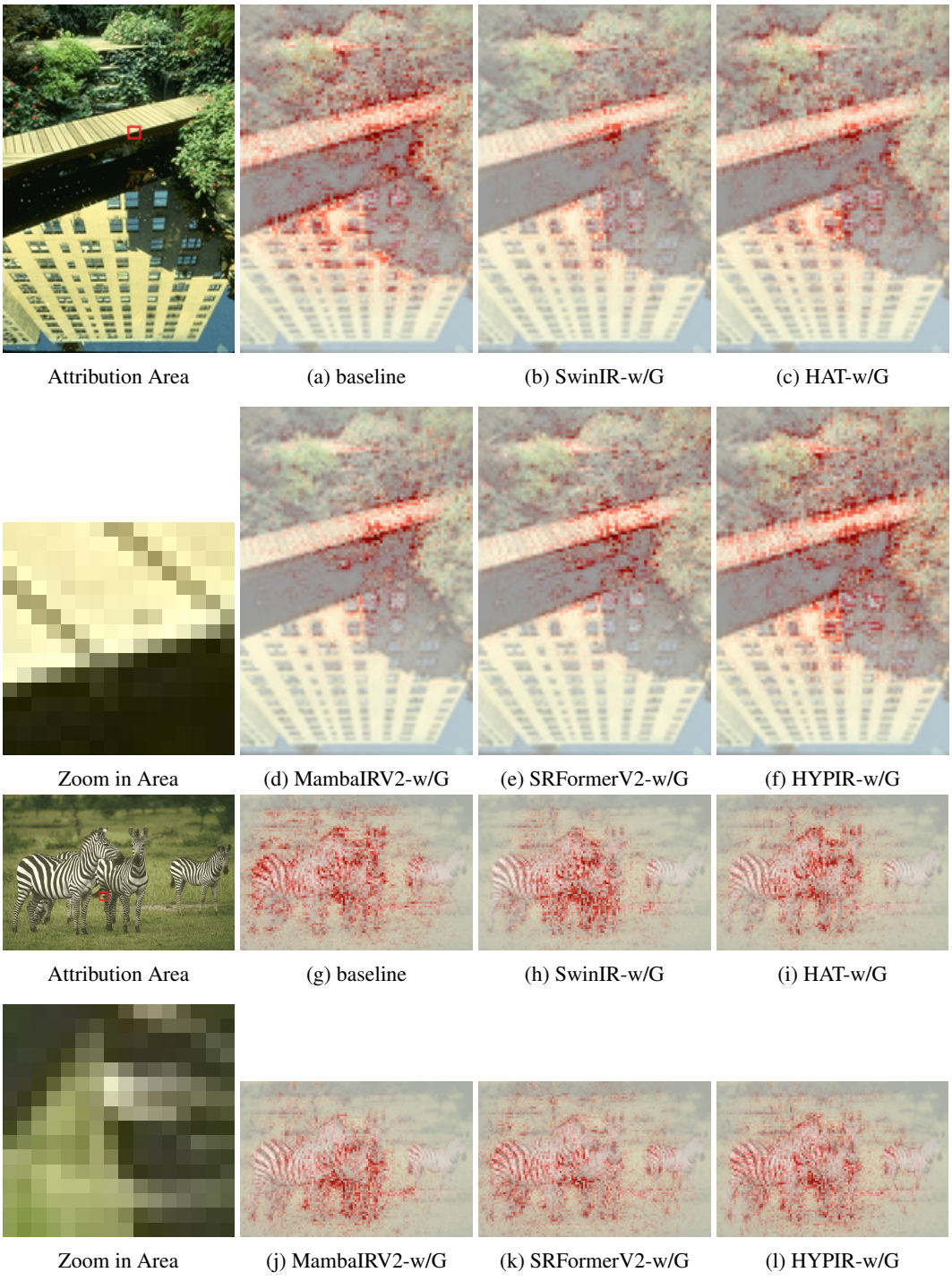


Figure 10: Local attribution maps at $4\times$.

$\mathcal{L}_{\text{irc}}$ complements the other terms as a degradation-cycle regularizer. It can improve perceptual scores with modest PSNR movement, but it does not replace $\mathcal{L}_{\text{isc}}$. Because the student architecture is unchanged in all configurations, these gains reflect the proposed ISCD supervision rather than additional inference-time capacity. Overall, the results verify the three-term design in Sec. 3.2 and show that ISCD improvements are not reducible to conventional output distillation.

Table 16: Progressive ISCD loss ablation with HAT teacher (five-benchmark mean). **Baseline**: $\mathcal{L}_{\mathrm{REC}}$ only (Table 3). All runs are ISCD-only without GAMD. **Full ISCD** denotes $\mathcal{L}_{\mathrm{out}}+\mathcal{L}_{\mathrm{irc}}+\mathcal{L}_{\mathrm{isc}}$. ↑ and ↓ indicate higher and lower is better, respectively. **Bold** indicates the best performance in ISCD ablation experiments.

Scale	Components	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
2×	Baseline	31.3700	0.9142	0.0996	31.0004	5.2242	0.6203	66.8814	0.4225
	$\mathcal{L}_{\mathrm{out}}$	31.8936	0.9204	0.0913	30.0189	5.1038	0.6367	67.3207	0.4354
	$\mathcal{L}_{\mathrm{out}}+\mathcal{L}_{\mathrm{irc}}$	31.9082	0.9207	0.0911	29.9181	5.1275	**0.6374**	67.3283	0.4352
	$\mathcal{L}_{\mathrm{out}}+\mathcal{L}_{\mathrm{isc}}$	31.9530	0.9212	0.0901	30.1934	5.1368	0.6331	67.3776	0.4365
	Full ISCD	**31.9655**	**0.9212**	**0.0877**	**29.8218**	**5.0656**	0.6351	**67.3883**	**0.4379**
4×	Baseline	25.5859	0.7408	0.3497	51.3503	8.0077	0.4265	44.9011	0.2767
	$\mathcal{L}_{\mathrm{out}}$	26.1827	0.7739	0.2908	48.0659	6.6917	0.4785	53.3278	0.3296
	$\mathcal{L}_{\mathrm{out}}+\mathcal{L}_{\mathrm{irc}}$	26.1836	0.7740	0.2901	48.2404	6.6738	0.4828	53.9576	0.3326
	$\mathcal{L}_{\mathrm{out}}+\mathcal{L}_{\mathrm{isc}}$	26.1882	0.7754	**0.2870**	47.9534	6.5600	0.4798	54.2863	0.3344
	Full ISCD	**26.1969**	**0.7754**	0.2881	**47.9308**	**6.5597**	**0.4840**	**54.4313**	**0.3354**

Table 17: Progressive ISCD loss ablation with HYPIR teacher (five-benchmark mean). **Baseline**: $\mathcal{L}_{\mathrm{REC}}$ only (Table 3). All runs are ISCD-only without GAMD. ↑ and ↓ indicate higher and lower is better, respectively. **Bold** indicates the best performance in ISCD ablation experiments.

Scale	Components	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
2×	Baseline	31.3700	0.9142	0.0996	31.0004	5.2242	0.6203	66.8814	0.4225
	$\mathcal{L}_{\mathrm{out}}$	30.8456	0.8986	0.1280	32.0777	5.3046	0.6019	66.6199	0.4195
	$\mathcal{L}_{\mathrm{out}}+\mathcal{L}_{\mathrm{irc}}$	31.0882	0.9112	0.1011	31.2472	5.3268	0.6162	66.8638	0.4168
	$\mathcal{L}_{\mathrm{out}}+\mathcal{L}_{\mathrm{isc}}$	31.7216	0.9172	0.0954	30.9702	5.0656	0.6236	67.2559	0.4310
	Full ISCD	**31.7675**	**0.9183**	**0.0928**	**30.3341**	**5.0543**	**0.6264**	**67.3133**	**0.4310**
4×	Baseline	25.5859	0.7408	0.3497	51.3503	8.0077	0.4265	44.9011	0.2767
	$\mathcal{L}_{\mathrm{out}}$	25.6216	0.7422	0.3517	51.4405	7.6412	0.4233	45.6980	0.2851
	$\mathcal{L}_{\mathrm{out}}+\mathcal{L}_{\mathrm{irc}}$	25.6128	0.7424	0.3507	52.0720	8.6297	0.4160	46.8415	0.2921
	$\mathcal{L}_{\mathrm{out}}+\mathcal{L}_{\mathrm{isc}}$	25.5727	0.7437	**0.3446**	**49.9643**	7.6121	0.4212	**48.3394**	**0.3039**
	Full ISCD	**25.6264**	**0.7447**	0.3455	50.0120	**7.5800**	**0.4266**	48.3149	0.3034

A.14 Additional controls for the ISCD inverse proxy

We further examine whether the proposed learned student-isomorphic inverse proxy can be replaced by alternative feature spaces. Four configurations are compared in Table 18. **RRDB-Inv** denotes our proposed inverse network, which is architecturally isomorphic to the RRDB student, pretrained on the HR-to-LR degradation task, and frozen during distillation. It is used by both the multi-layer contrastive loss $\mathcal{L}_{\mathrm{isc}}$ and the inverse reconstruction loss $\mathcal{L}_{\mathrm{irc}}$.

Iso-Rand retains the same architecture as the student but uses randomly initialized frozen parameters. Its intermediate features are used by $\mathcal{L}_{\mathrm{isc}}$. Because its final output does not provide a meaningful LR reconstruction, $\mathcal{L}_{\mathrm{irc}}$ is disabled. This configuration evaluates whether architectural isomorphism and fixed random features alone can reproduce the gains of complete ISCD. Since it does not use $\mathcal{L}_{\mathrm{irc}}$, it is a stress-test baseline rather than a factorial isolation of degradation pretraining.

EDSR-Inv replaces RRDB-Inv with an EDSR network pretrained on the same HR-to-LR degradation task. The network is frozen and used with the complete ISCD objective. This comparison retains task-specific inverse pretraining and cycle supervision while removing architectural correspondence with the RRDB student. It therefore evaluates whether a heterogeneous degradation network can provide an equally effective proxy space.

VGG uses an ImageNet-pretrained VGG extractor for multi-layer contrastive alignment, following the protocol in Appendix A.16. Because VGG does not implement an HR-to-LR inverse mapping, $\mathcal{L}_{\mathrm{irc}}$ is disabled in this configuration. It represents a generic perceptual feature space that is neither architecturally aligned with the student nor pretrained on the SR degradation task. The VGG control therefore evaluates a generic perceptual alignment alternative rather than a complete inverse-network replacement with identical loss components.

RRDB-Inv and EDSR-Inv provide the most direct comparison because both support the complete ISCD objective, while differing in architectural correspondence with the student. Iso-Rand and VGG

serve as broader controls for fixed random and generic perceptual representations, respectively, but do not support a meaningful inverse-reconstruction constraint. The results therefore support the complete RRDB-Inv design, while the RRDB-Inv versus EDSR-Inv comparison more specifically isolates the benefit of student–inverse architectural correspondence.

Table 18: Controls for the ISCD inverse proxy. **Iso-Rand**: frozen randomly initialized student-isomorphic proxy without $\mathcal{L}_{\mathrm{irc}}$. **EDSR-Inv**: degradation-pretrained heterogeneous EDSR (Lim et al., 2017a) inverse network. **VGG**: ImageNet-pretrained VGG alignment with $\mathcal{L}_{\mathrm{irc}}$ disabled (reproduced from Table 20). **RRDB-Inv**: the proposed degradation-pretrained student-isomorphic inverse network("-ISCD" in Table 20). ↑/↓: higher/lower better. **Bold** values indicate the best result at each scale.

Scale	Proxy	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
SwinIR 3×	Iso-Rand	28.1598	0.8365	0.2090	41.3453	5.9902	0.5449	61.0147	0.3762
	EDSR-Inv	28.1597	0.8377	0.1990	40.8060	5.7437	0.5512	62.5243	0.3927
	VGG	28.1644	0.8379	0.2034	40.9588	5.8308	0.5495	61.8905	0.3863
	RRDB-Inv	**28.1699**	**0.8424**	**0.1926**	**40.2894**	**5.6425**	**0.5672**	**63.5977**	**0.4095**
HYPIR 4×	Iso-Rand	**25.6385**	0.7423	0.3596	52.8037	8.6931	0.4267	44.0291	0.2731
	EDSR-Inv	25.5008	0.7413	0.3472	50.3299	7.7002	0.4308	46.9357	0.2922
	VGG	25.6242	0.7444	0.3463	51.4473	**7.4950**	**0.4317**	46.4123	0.2884
	RRDB-Inv	25.6264	**0.7447**	**0.3455**	**50.0120**	7.5800	0.4266	**48.3149**	**0.3034**

Comparison with the randomly initialized isomorphic proxy. With SwinIR at $3\times$, RRDB-Inv outperforms Iso-Rand on all eight metrics. It reduces LPIPS from 0.2090 to 0.1926, improves MUSIQ from 61.01 to 63.60, and raises MANIQA from 0.3762 to 0.4095. With HYPIR at $4\times$, RRDB-Inv leads on six of eight metrics. Iso-Rand attains a marginal 0.012 dB PSNR advantage and a negligible 0.0001 CLIPIQA advantage, whereas RRDB-Inv substantially improves LPIPS, BRISQUE, NIQE, MUSIQ, and MANIQA. These results show that architectural isomorphism and fixed random features alone are insufficient to reproduce the overall performance of complete ISCD. Because Iso-Rand does not use $\mathcal{L}_{\mathrm{irc}}$, this comparison evaluates the complete learned inverse design rather than isolating degradation pretraining alone.

Comparison with the heterogeneous EDSR inverse network. RRDB-Inv outperforms EDSR-Inv on all eight metrics with SwinIR at $3\times$, including a 1.07 gain in MUSIQ and a 0.0168 gain in MANIQA. With HYPIR at $4\times$, RRDB-Inv leads on seven of eight metrics, improving PSNR by 0.13 dB and MUSIQ by 1.38. EDSR-Inv is better only on CLIPIQA. Since both inverse networks are pretrained on the same HR-to-LR task and retain the complete ISCD objective, this comparison supports the benefit of architectural correspondence between the student and inverse network.

Comparison with the VGG proxy. At $3\times$, RRDB-Inv outperforms VGG on all eight metrics, reducing LPIPS from 0.2034 to 0.1926 and improving MUSIQ from 61.89 to 63.60. At $4\times$, RRDB-Inv leads VGG on six of eight metrics. VGG performs better on NIQE and CLIPIQA, whereas RRDB-Inv provides stronger fidelity and learned perceptual quality, including improvements in PSNR, SSIM, LPIPS, MUSIQ, and MANIQA. This comparison indicates that generic classification features cannot provide the same degradation-aware alignment as the proposed inverse proxy.

Summary. RRDB-Inv is best on all eight metrics with SwinIR at $3\times$ and on five of eight metrics with HYPIR at $4\times$. Its remaining exceptions are a marginal PSNR advantage for Iso-Rand and better NIQE and CLIPIQA under VGG. Overall, the results consistently favor the proposed proxy on full-reference fidelity and learned perceptual metrics. They demonstrate that the combination of learned inverse representations and student-inverse architectural correspondence provides more effective heterogeneous supervision than random isomorphic features, a degradation-pretrained heterogeneous network, or a generic perceptual extractor.

A.15 Single-layer versus multi-layer inverse alignment

ISCD applies contrastive alignment to representations extracted at multiple depths of the frozen inverse network. We examine whether these features provide complementary supervision by comparing the complete multi-layer objective with a single-layer variant. The single-layer variant uses

the deepest intermediate representation of the inverse network, extracted after the second scale-conversion block and immediately before the reconstruction head that projects latent features to the three-channel LR output. This layer provides a strong single-feature baseline because it is the closest intermediate representation to the final inverse reconstruction.

Let $\ell_{\mathrm{con}}(\mathbf{Z}^S, \mathbf{Z}^T)$ denote the batch-wise contrastive objective

$$\ell_{\mathrm{con}}(\mathbf{Z}^S, \mathbf{Z}^T) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big(\kappa(\mathbf{z}_i^S)^\top \mathbf{z}_i^T\big)}{\sum_{j=1}^{B} \exp\big(\kappa(\mathbf{z}_i^S)^\top \mathbf{z}_j^T\big)}, \tag{18}$$

where $\mathbf{z}_i^S$ and $\mathbf{z}_i^T$ are flattened and ℓ_2-normalized inverse features, and $\kappa = 0.2$ is the logit scale. The single-layer objective is

$$\mathcal{L}_{\mathrm{isc}}^{\mathrm{single}} = W\, \ell_{\mathrm{con}}\left(\mathbf{Z}^{S,(l_{\mathrm{deep}})}, \mathbf{Z}^{T,(l_{\mathrm{deep}})}\right), \tag{19}$$

where l_{deep} is the deepest exported intermediate layer and $W = \sum_{l\in\mathbb{L}} w_l$. This preserves the aggregate contrastive weight of the multi-layer configuration.

Both variants retain the same $\mathcal{L}_{\mathrm{out}}$, $\mathcal{L}_{\mathrm{irc}}$, and calibrated loss coefficients. The aggregate weight assigned to contrastive alignment is also held constant. Thus, the comparison changes only whether inverse supervision is drawn from one depth or multiple depths.

Table 19: Single-intermediate-layer versus multi-layer ISCD alignment with SwinIR at $2\times$. Both configurations use the same aggregate contrastive weight. ↑/↓: higher/lower better. **Bold** values indicate the better result.

Alignment	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
Single layer	31.9056	0.9205	0.0916	30.8306	5.1084	0.6317	67.1843	0.4367
Multi-layer	**31.9716**	**0.9214**	**0.0866**	**29.5424**	**5.0599**	**0.6327**	**67.6019**	**0.4412**

As shown in Table 19, Multi-layer alignment outperforms the single-layer variant on all eight metrics. It increases PSNR from 31.9056 to 31.9716 dB and reduces LPIPS from 0.0916 to 0.0866. It also lowers BRISQUE by 1.29 and improves MUSIQ from 67.18 to 67.60. Although the improvements in SSIM and CLIPIQA are modest, their directions remain consistent with the other metrics. These results show that the deepest intermediate feature alone does not capture all useful inverse representations and that aligning features across multiple depths provides complementary fidelity and perceptual supervision.

A.16 Student-isomorphic inverse alignment versus VGG-based feature matching

Objective. The preceding experiments show that multi-layer contrastive alignment is the primary source of teacher-dependent gains in ISCD. We next examine whether these gains can be reproduced by a generic perceptual feature extractor. Specifically, we compare the proposed student-isomorphic inverse network with an ImageNet-pretrained VGG backbone across five heterogeneous teachers and three upscaling factors.

Comparison protocol. Both configurations use the same RRDB student, teacher reconstructions, DF2K training data, optimization schedule, reconstruction loss, output-distillation loss, and batch-wise contrastive formulation. They differ in the proxy used for contrastive alignment. The **VGG** configuration extracts multi-layer features from a frozen ImageNet-pretrained VGG network. Because VGG does not implement an HR-to-LR mapping, the inverse reconstruction loss $\mathcal{L}_{\mathrm{irc}}$ is disabled. The **ISCD** configuration extracts features from the frozen student-isomorphic inverse network and retains $\mathcal{L}_{\mathrm{irc}}$.

This comparison evaluates the complete applicable design of each proxy. It is not a factorial isolation of the feature extractor because only the inverse network supports the inverse reconstruction constraint. Nevertheless, the loss-component ablation in Appendix A.13 separately shows that $\mathcal{L}_{\mathrm{isc}}$ is the primary driver of teacher-dependent gains, while $\mathcal{L}_{\mathrm{irc}}$ provides complementary regularization. The RRDB-Inv versus EDSR-Inv control in Appendix A.14 further isolates architectural correspondence while retaining the complete three-term objective.

Table 20: Comparison between VGG-based feature matching and the proposed ISCD design over five benchmark datasets. **Baseline**: RRDB trained with $\mathcal{L}_{\mathrm{REC}}$ only. **VGG**: output distillation and VGG-based multi-layer contrastive alignment, with $\mathcal{L}_{\mathrm{irc}}$ disabled. **ISCD**: output distillation, multi-layer contrastive alignment through the frozen student-isomorphic inverse network, and inverse reconstruction. ↑/↓: higher/lower better. **Bold** values indicate the better result between VGG and ISCD for the same teacher and scale.

Scale	Configuration	PSNR↑	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
	baseline	31.3700	0.9142	0.0996	31.0004	5.2242	0.6203	66.8814	0.4225
	SwinIR-vgg	31.9205	0.9206	0.0932	30.0226	5.1525	**0.6359**	67.2252	0.4340
	SwinIR-ISCD	**31.9716**	**0.9214**	**0.0866**	**29.5424**	**5.0599**	0.6327	**67.6019**	**0.4412**
	HAT-vgg	31.9280	0.9207	0.0927	30.4551	5.1491	**0.6389**	67.2049	0.4335
	HAT-ISCD	**31.9655**	**0.9212**	**0.0877**	**29.8218**	**5.0656**	0.6351	**67.3883**	**0.4379**
2×	MambaIRV2-vgg	31.9162	0.9204	0.0940	30.1305	5.1385	0.6371	67.0905	0.4315
	MambaIRV2-ISCD	**31.9557**	**0.9210**	**0.0886**	**30.1076**	**5.0813**	**0.6384**	**67.2937**	**0.4369**
	SRFormerV2-vgg	31.9073	0.9205	0.0930	30.1387	5.2179	**0.6357**	67.2481	0.4337
	SRFormerV2-ISCD	**31.9400**	**0.9210**	**0.0874**	**29.6413**	**5.0642**	0.6345	**67.4953**	**0.4402**
	HYPIR-vgg	30.8950	0.9014	0.1248	32.2874	5.2654	0.6048	66.8872	0.4212
	HYPIR-ISCD	**31.7675**	**0.9183**	**0.0928**	**30.3341**	**5.0543**	**0.6264**	**67.3133**	**0.4310**
	baseline	27.6764	0.8170	0.2493	44.5242	6.7100	0.4994	54.8755	0.3314
	SwinIR-vgg	28.1644	0.8379	0.2034	40.9588	5.8308	0.5495	61.8905	0.3863
	SwinIR-ISCD	**28.1699**	**0.8424**	**0.1926**	**40.2894**	**5.6425**	**0.5672**	**63.5977**	**0.4095**
	HAT-vgg	28.1273	0.8366	0.2040	40.9471	5.8379	0.5511	61.6059	0.3805
	HAT-ISCD	**28.1285**	**0.8407**	**0.1976**	**40.7580**	**5.7502**	**0.5648**	**62.6522**	**0.3958**
3×	MambaIRV2-vgg	28.1094	0.8361	0.2078	41.9509	5.9611	0.5499	60.8828	0.3746
	MambaIRV2-ISCD	**28.1173**	**0.8400**	**0.2011**	**41.6866**	**5.8166**	**0.5662**	**62.4254**	**0.3933**
	SRFormerV2-vgg	28.1320	0.8369	0.2055	41.2665	5.9317	0.5490	61.2313	0.3767
	SRFormerV2-ISCD	**28.1363**	**0.8411**	**0.1954**	**40.5316**	**5.6650**	**0.5690**	**63.1487**	**0.4014**
	HYPIR-vgg	27.5164	0.8095	0.2647	45.7784	6.8438	0.4917	55.4145	0.3359
	HYPIR-ISCD	**27.6876**	**0.8176**	**0.2449**	**43.9110**	**6.6357**	**0.5002**	**55.7559**	**0.3360**
	baseline	25.5859	0.7408	0.3497	51.3503	8.0077	0.4265	44.9011	0.2767
	SwinIR-vgg	**26.2703**	0.7774	0.2815	47.6517	6.3948	0.4820	55.8783	0.3437
	SwinIR-ISCD	26.2567	**0.7833**	**0.2688**	**45.0991**	**5.9216**	**0.5143**	**59.6175**	**0.3815**
	HAT-vgg	26.1958	0.7745	0.2910	49.0493	6.6691	0.4755	53.7440	0.3288
	HAT-ISCD	**26.1969**	**0.7754**	**0.2881**	**47.9308**	**6.5597**	**0.4840**	**54.4313**	**0.3354**
4×	MambaIRV2-vgg	26.2266	0.7740	0.3020	49.2191	6.5918	0.4833	51.4718	0.3170
	MambaIRV2-ISCD	**26.2353**	**0.7821**	**0.2779**	**45.7632**	**6.0993**	**0.5126**	**57.7399**	**0.3594**
	SRFormerV2-vgg	26.2168	0.7766	0.2865	47.5131	6.4824	0.4898	54.9594	0.3378
	SRFormerV2-ISCD	**26.2173**	**0.7826**	**0.2697**	**45.3595**	**5.9526**	**0.5211**	**59.7020**	**0.3778**
	HYPIR-vgg	25.6242	0.7444	0.3463	51.4473	7.4950	**0.4317**	46.4123	0.2884
	HYPIR-ISCD	**25.6264**	**0.7447**	**0.3455**	**50.0120**	**7.5800**	0.4266	**48.3149**	**0.3034**

VGG as a control VGG-based supervision is widely used to compare reconstructed images in a fixed perceptual feature space. It avoids direct correspondence between heterogeneous teacher and student hidden states, making it a natural output-mediated baseline for ISCD. However, VGG is trained for image classification rather than HR-to-LR reconstruction. Its feature hierarchy therefore emphasizes generic semantic and perceptual information without explicitly modeling how an HR hypothesis maps back to the observed LR input. The proposed inverse network differs in two respects. First, it is trained on the inverse HR-to-LR task and therefore supplies representations directly associated with the reconstruction pipeline. Second, its backbone follows the student architecture, allowing features at multiple depths to provide structurally compatible supervision. These properties permit ISCD to combine contrastive feature alignment with inverse reconstruction while keeping the deployed student unchanged.

Overall quantitative comparison. Table 20 shows a consistent advantage for the complete ISCD design. Across five teachers, three scales, and eight metrics, ISCD outperforms VGG in 114 of the 120 teacher–scale–metric comparisons. On the five central fidelity and perceptual metrics, PSNR, SSIM, LPIPS, MUSIQ, and MANIQA, ISCD wins 74 of 75 comparisons. The only exception is SwinIR PSNR at $4\times$, where VGG is higher by 0.0136 dB.

At $2\times$, ISCD outperforms VGG for every teacher on PSNR, SSIM, LPIPS, BRISQUE, NIQE, MUSIQ, and MANIQA. The only exceptions occur on CLIPIQA for SwinIR, HAT, and SRFormerV2. The improvements are moderate for conventional high-fidelity teachers but remain directionally consistent. For example, SwinIR-ISCD improves PSNR from 31.9205 to 31.9716 dB, reduces LPIPS from 0.0932 to 0.0866, and increases MUSIQ from 67.2252 to 67.6019. The largest $2\times$ difference occurs with HYPIR. VGG-based alignment yields 30.8950 dB PSNR and 0.1248 LPIPS, falling below the undistilled student on both metrics. ISCD instead reaches 31.7675 dB and 0.0928 LPIPS while also improving all five no-reference metrics. This result is important because HYPIR differs from the CNN student in both architecture and reconstruction behavior. Generic VGG features do not provide reliable supervision in this case, whereas the inverse-network design retains positive transfer.

At $3\times$, ISCD outperforms VGG on all 40 reported comparisons. Fidelity gains are sometimes small, such as 0.0012 dB for HAT and 0.0043 dB for SRFormerV2, but perceptual improvements are substantially larger. With SwinIR, ISCD reduces LPIPS from 0.2034 to 0.1926 and raises MUSIQ from 61.8905 to 63.5977. With SRFormerV2, it reduces LPIPS from 0.2055 to 0.1954 and raises MANIQA from 0.3767 to 0.4014. The consistency across all teachers and metrics indicates that the advantage is not restricted to a particular teacher family.

At $4\times$, ISCD wins 37 of 40 comparisons. It outperforms VGG on SSIM, LPIPS, BRISQUE, MUSIQ, and MANIQA for every teacher. The remaining exceptions are limited to SwinIR PSNR and HYPIR NIQE and CLIPIQA. For SwinIR, the 0.0136 dB PSNR difference is accompanied by a considerably stronger perceptual result under ISCD, including LPIPS of 0.2688 versus 0.2815 and MUSIQ of 59.6175 versus 55.8783. MambaIRV2 exhibits the same pattern more strongly, with similar PSNR but an LPIPS reduction from 0.3020 to 0.2779 and a MUSIQ increase from 51.4718 to 57.7399.

Fidelity and perceptual behavior. Fig. 11 summarizes the $2\times$ results across all eight metrics. ISCD improves full-reference fidelity and learned perceptual metrics together for every teacher. This differs from the common perception–distortion trade-off induced by generic perceptual supervision. At $2\times$ and $3\times$, the inverse-network design improves PSNR, SSIM, LPIPS, MUSIQ, and MANIQA simultaneously across all teacher architectures. At $4\times$, where the inverse problem is more difficult, ISCD still preserves comparable PSNR while providing substantially stronger perceptual performance.

These results suggest that the observed trade-off depends partly on the supervision space. VGG features emphasize generic semantic similarity, whereas the inverse network organizes supervision around the HR-to-LR reconstruction relation. The latter provides a stronger connection between perceptual structure and input fidelity, reducing the need to improve one at the expense of the other. This interpretation is supported by the proxy controls in Appendix A.14, where RRDB-Inv also outperforms the degradation-pretrained EDSR inverse proxy under the same complete objective.

Role of inverse reconstruction. The VGG control does not include $\mathcal{L}_{\mathrm{irc}}$, so Table 20 should be interpreted as a comparison between complete applicable designs rather than as an extractor-only ablation. However, the advantage of ISCD cannot be attributed solely to the presence of inverse reconstruction. Appendix A.13 shows that adding $\mathcal{L}_{\mathrm{isc}}$ to output distillation produces the dominant teacher-dependent improvements, while $\mathcal{L}_{\mathrm{irc}}$ generally provides smaller complementary gains. Moreover, the EDSR-Inv control in Appendix A.14 retains $\mathcal{L}_{\mathrm{irc}}$ but remains consistently weaker than the student-isomorphic RRDB-Inv proxy. Together, these results support both components of the design: task-specific inverse representations provide the principal alignment benefit, and inverse reconstruction supplies an additional consistency constraint.

Robustness across teacher architectures. The inverse-network design also reduces sensitivity to teacher architecture. At $2\times$, the PSNR of the five teachers spans 25.6756 to 33.6690 dB, a range of 7.9934 dB. The corresponding VGG-distilled students span 30.8950 to 31.9280 dB, a range of 1.0330 dB, with the HYPIR student remaining a clear outlier. ISCD-distilled students instead span only 31.7675 to 31.9716 dB, reducing the range to 0.2041 dB while keeping every student above the undistilled baseline.

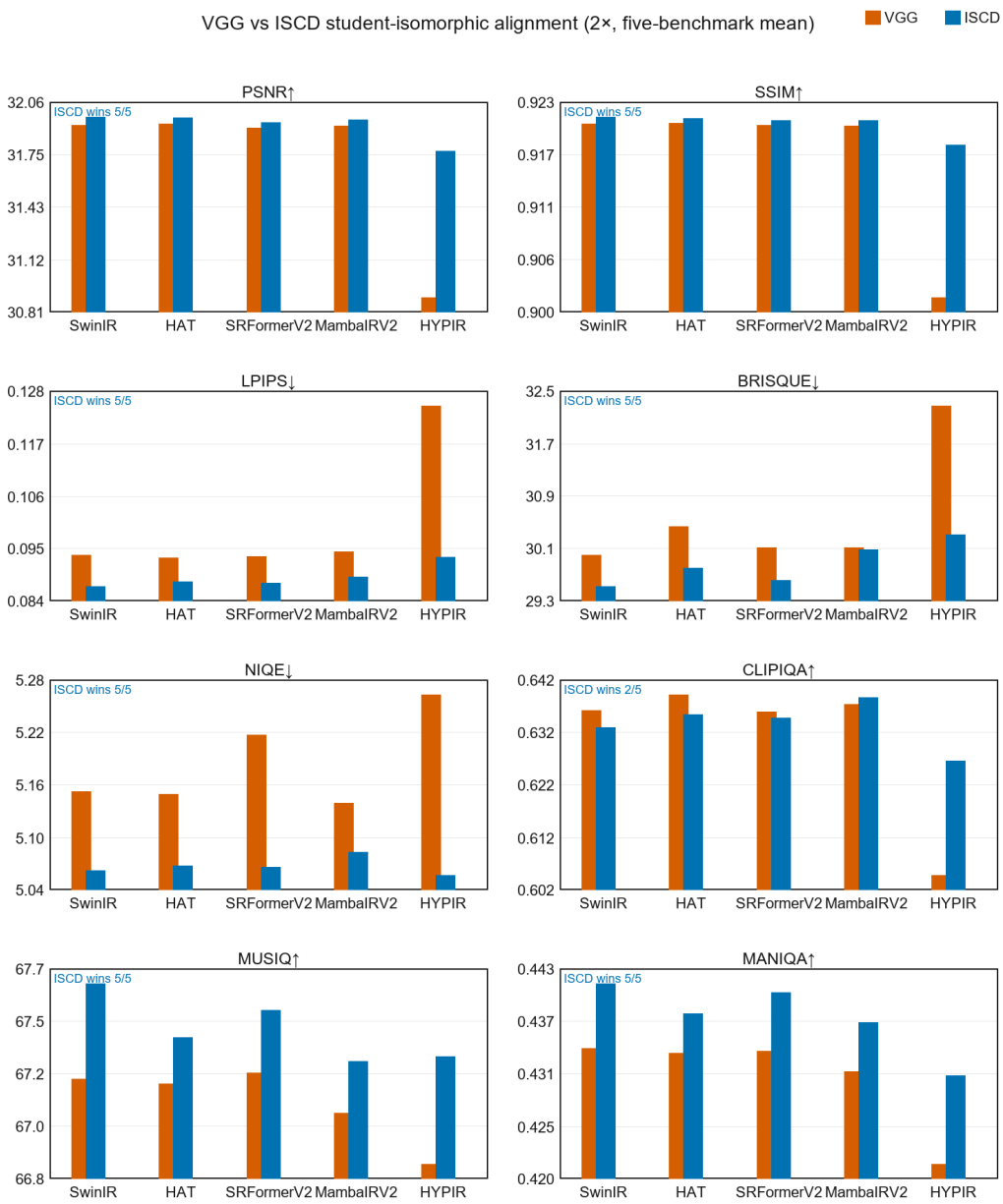


Figure 11: VGG versus ISCD at 2× over five benchmarks. Each panel reports one metric and the number of teachers for which ISCD performs better. ISCD wins all teacher comparisons on seven of eight metrics, with isolated VGG advantages on CLIPIQA.

Fig. 12 visualizes this behavior across all metrics. The tight clustering of ISCD students does not imply that teacher identity is irrelevant. Rather, it shows that the shared inverse proxy limits architecture-specific negative transfer while retaining useful teacher-dependent supervision. The HYPIR result is particularly informative because ISCD converts a low-fidelity diffusion teacher into positive student transfer without directly matching its hidden representation.

Relation to CSD-style supervision. The VGG configuration follows the output-mediated principle used by CSD-style methods. Both teacher and student reconstructions are mapped into a fixed external feature space, avoiding direct access to teacher hidden states. Table 20 shows that retaining

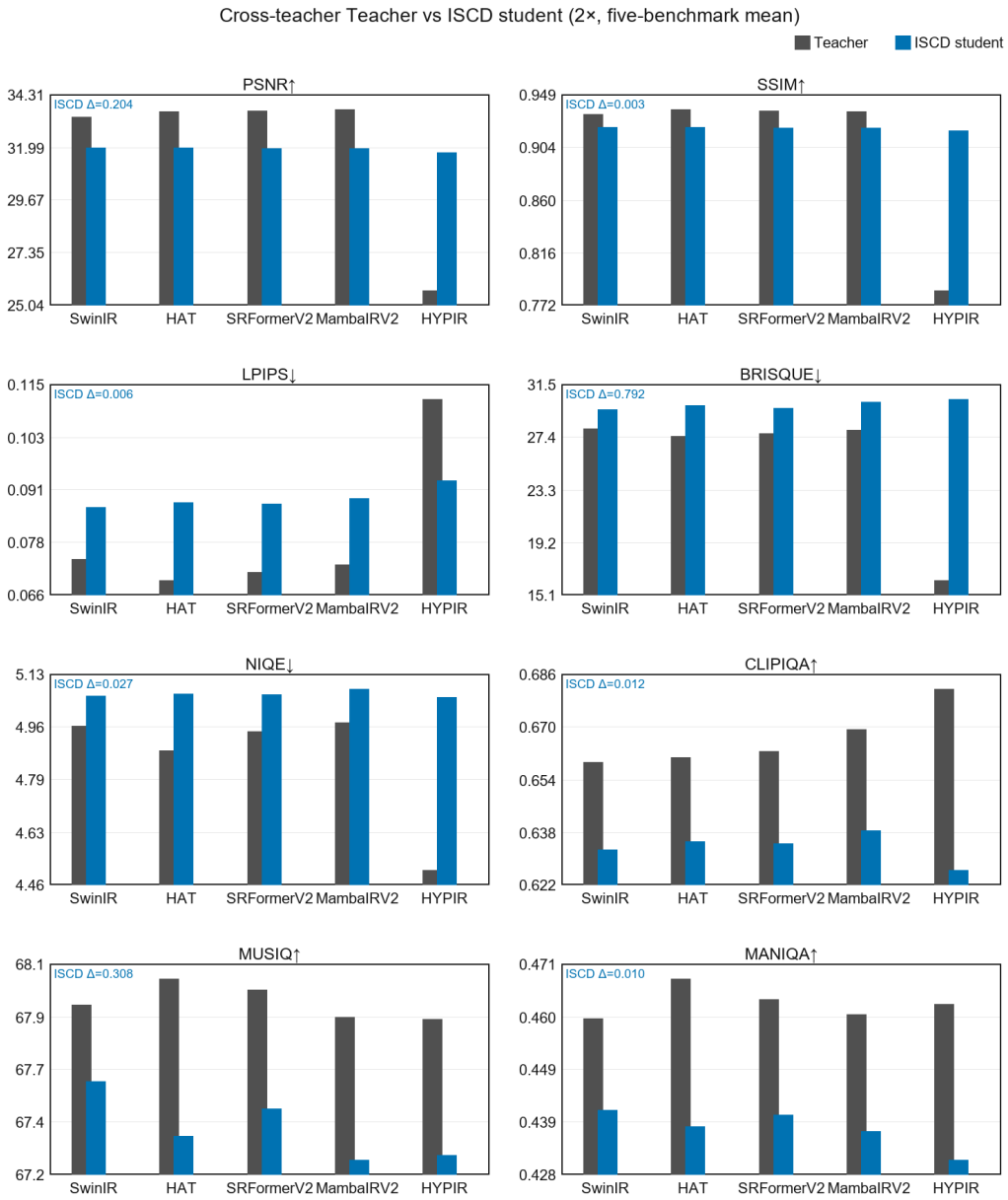


Figure 12: Cross-teacher robustness at $2\times$. Gray bars denote heterogeneous teachers and blue bars denote the corresponding ISCD-distilled students. Despite substantial variation among teacher architectures and fidelity levels, ISCD students exhibit a considerably narrower performance range, particularly in the low-fidelity HYPIR setting.

the same contrastive formulation is insufficient when the proxy lacks task-specific inverse structure. ISCD improves upon this strategy by replacing the generic classification feature space with a frozen HR-to-LR proxy whose backbone corresponds to the student.

Scope of the evidence. The VGG comparison establishes that the complete inverse-network design is substantially stronger than a generic perceptual alternative. It does not by itself isolate architectural isomorphism, inverse-task pretraining, and $\mathcal{L}_{\mathrm{irc}}$. These factors are separated by the complementary experiments in Appendix A.13, Appendix A.14, and Appendix A.15. Taken together,

the four studies provide converging evidence for the proposed design: contrastive inverse alignment drives teacher-dependent transfer, inverse reconstruction improves consistency, student-isomorphic pretraining strengthens the proxy space, and multi-layer features provide complementary supervision.

Summary. Across 120 pairwise metric comparisons, ISCD outperforms VGG in 114 cases and achieves a clean sweep at $3\times$. Its advantage is especially strong on LPIPS, MUSIQ, and MANIQA, while PSNR remains higher or nearly identical in all but one configuration. Combined with the loss-component, proxy, and layer-depth controls, these results support the complete ISCD design as a more effective and architecture-robust alternative to VGG-based feature matching for heterogeneous SR distillation.

A.17 Additional visual results of AMID

We provide additional qualitative comparisons on diverse challenging scenes, including urban structures, natural landscapes, and text-rich images. As illustrated in Fig 13, Fig 14 and Fig 15, AMID consistently reconstructs sharper boundaries and more faithful structural details compared to KD, CSD, AugKD, and DCKD, particularly in regions with dense high-frequency patterns. These visual trends remain stable across all evaluated scale factors, demonstrating that the behavioral-pattern alignment and degradation-cycle consistency generalize robustly beyond the quantitative benchmarks.

A.18 Limitations

Our current study has several limitations that point to avenues for future work.

Computational overhead of multi-step attribution. GAMD relies on Integrated Gradients to compute exact pixel-wise attributions, which requires multiple forward and backward passes during training. This introduces additional computational overhead compared to standard output-level distillation methods such as KD. However, this cost is inherent to any gradient-based attribution technique and is incurred only during the student's training phase, leaving inference completely unaffected. The overhead can be substantially reduced by adopting approximation methods with fewer integration steps or by incorporating fast attribution surrogates, which we leave for future investigation.

Dependence on a matched inverse model. ISCD requires a pretrained inverse network that follows the same HR-to-LR pairing used for student training. The inverse therefore stays inside the training pair distribution and does not estimate per-image degradation parameters at test time. This match need not hold for real-camera observations whose imaging statistics leave that pair distribution. Evaluating and adapting AMID under such observations remains future work.

Dependency on gradient-based attribution. GAMD formulates knowledge transfer entirely through attribution maps computed via path-integrated gradients, inheriting the general properties and limitations of attribution-based analysis. Attribution maps can be sensitive to the choice of baseline and interpolation path, although we follow the established practices in the literature and observe stable behavior across our experiments. The method does not address the complementary challenge of learning from teachers whose outputs are available only as discretely sampled images without gradient access. These aspects are inherent to the attribution paradigm itself and are shared by all methods that employ gradient-based interpretability tools.

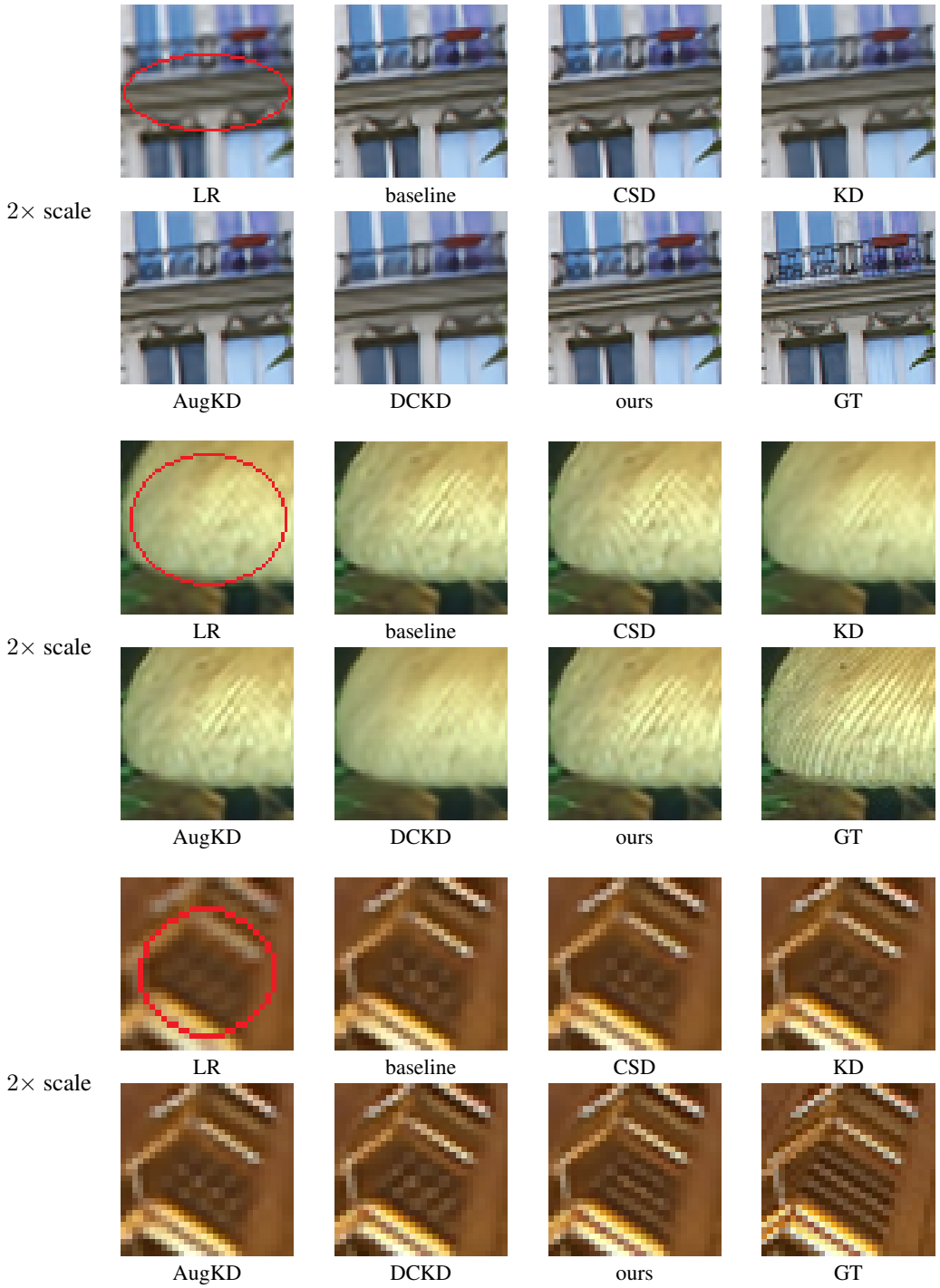


Figure 13: 2× qualitative comparison. **LR**: low-resolution input; **GT**: ground-truth. **baseline**: undistilled student. Columns labeled **CSD** (Wang et al., 2021b), **KD** (Hinton et al., 2015), **AugKD** (Zhang et al., 2025), and **DCKD** (Zhou et al., 2024a) show the student outputs distilled with the corresponding methods. **ours**: GAMD + ISCD. Fine-grained differences are best viewed when zoomed in.

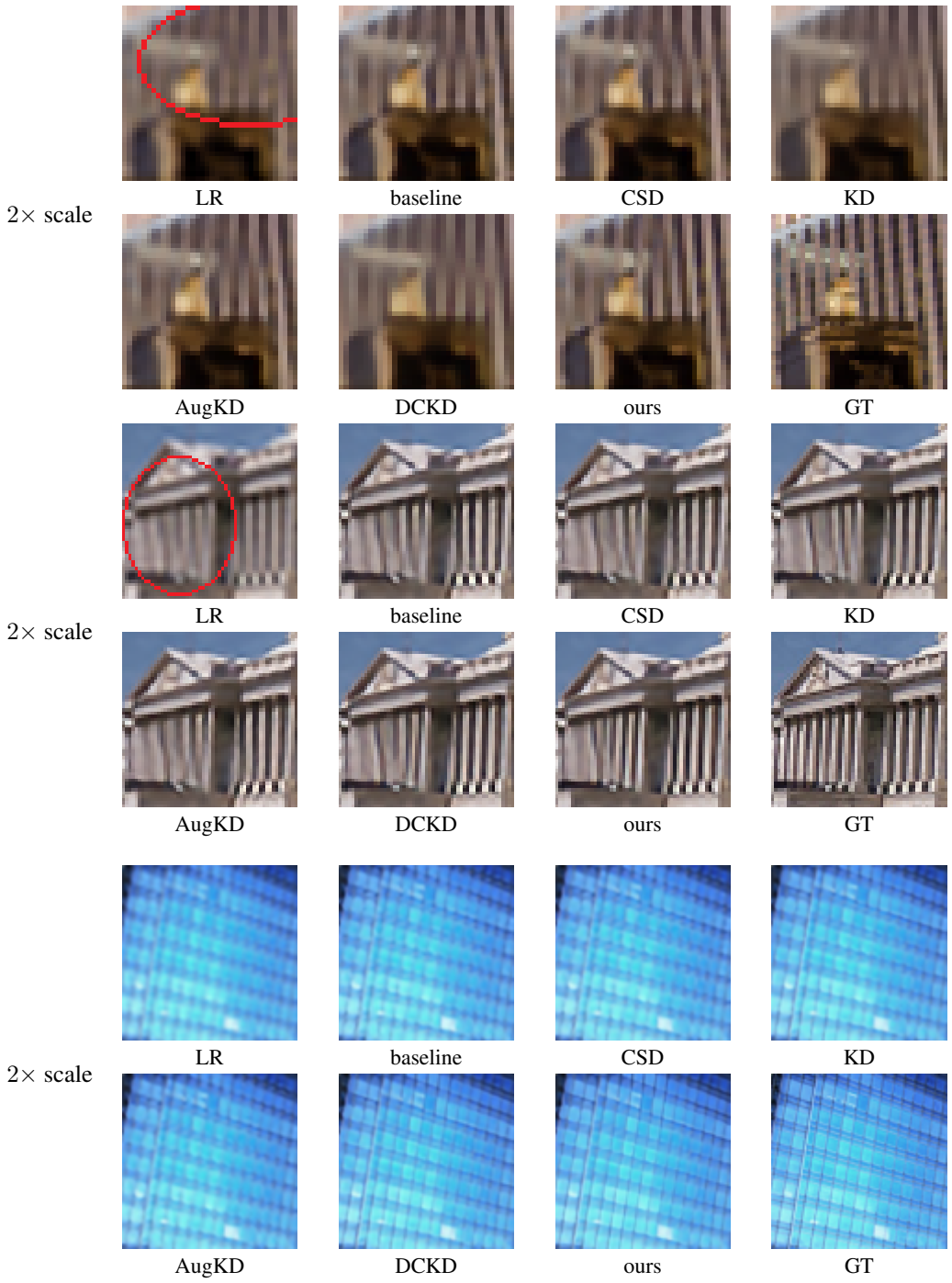


Figure 13: 2× qualitative comparison (continued).

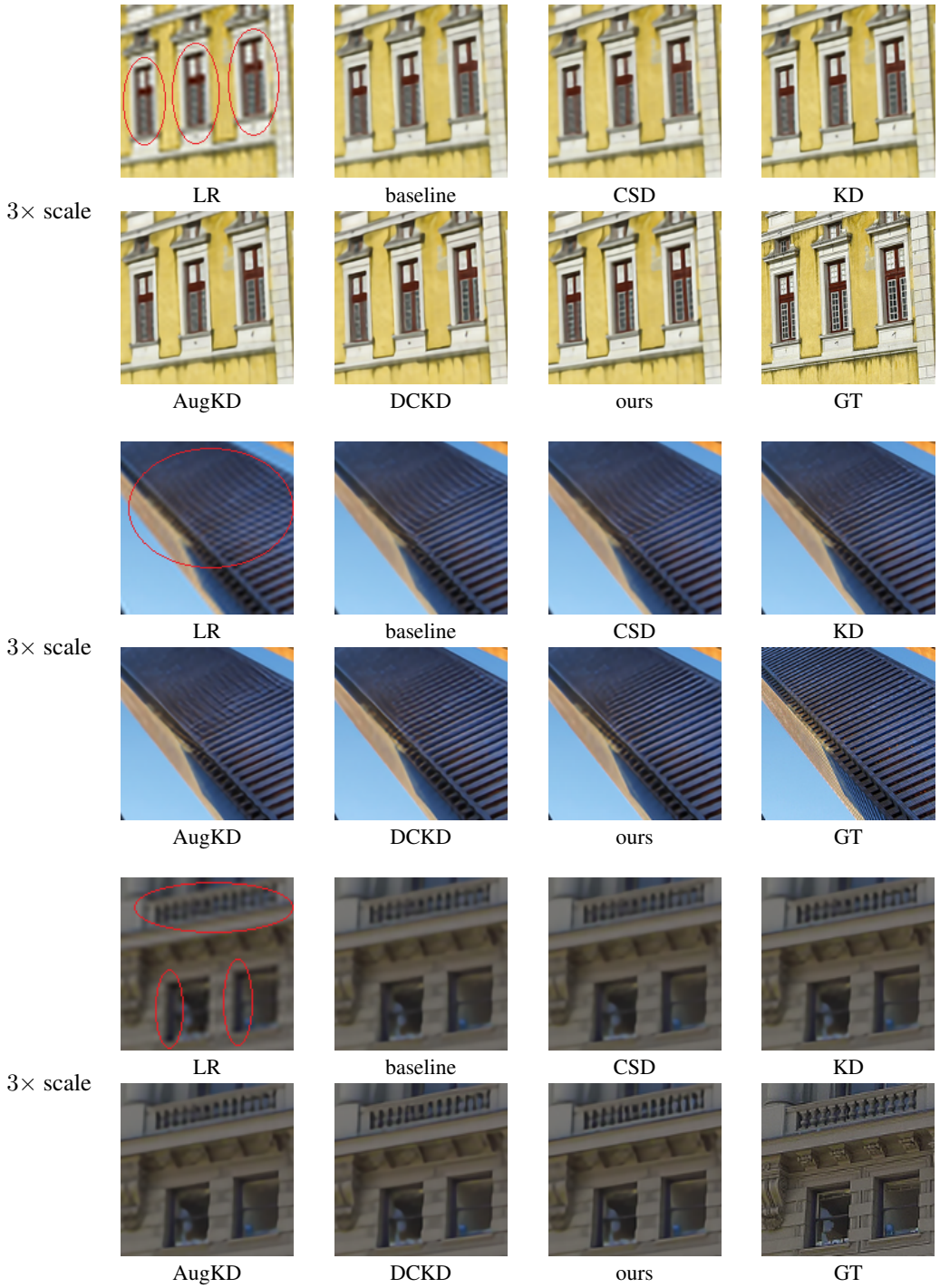


Figure 14: 3× qualitative comparison. **LR**: low-resolution input; **GT**: ground-truth. **baseline**: undistilled student. Columns labeled **CSD** (Wang et al., 2021b), **KD** (Hinton et al., 2015), **AugKD** (Zhang et al., 2025), and **DCKD** (Zhou et al., 2024a) show the student outputs distilled with the corresponding methods. **ours**: GAMD + ISCD. Fine-grained differences are best viewed when zoomed in.

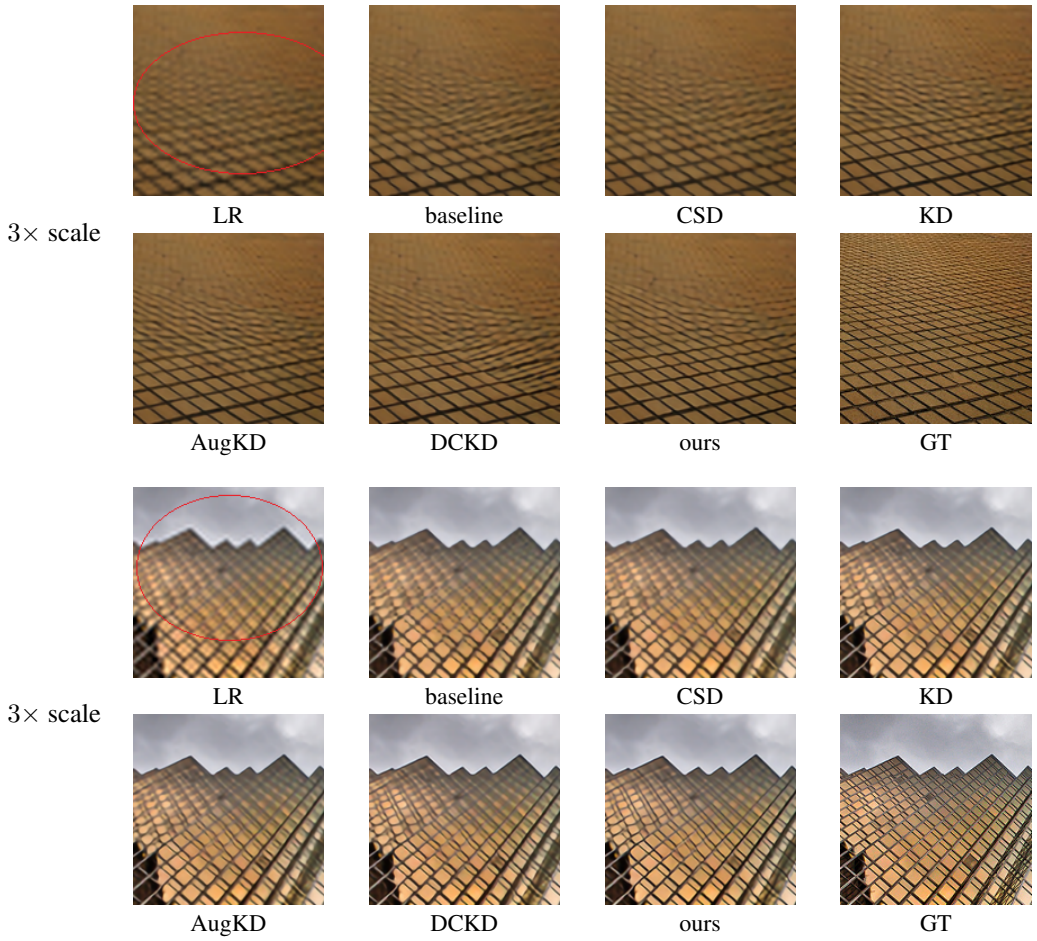


Figure 14: 3× qualitative comparison (continued).

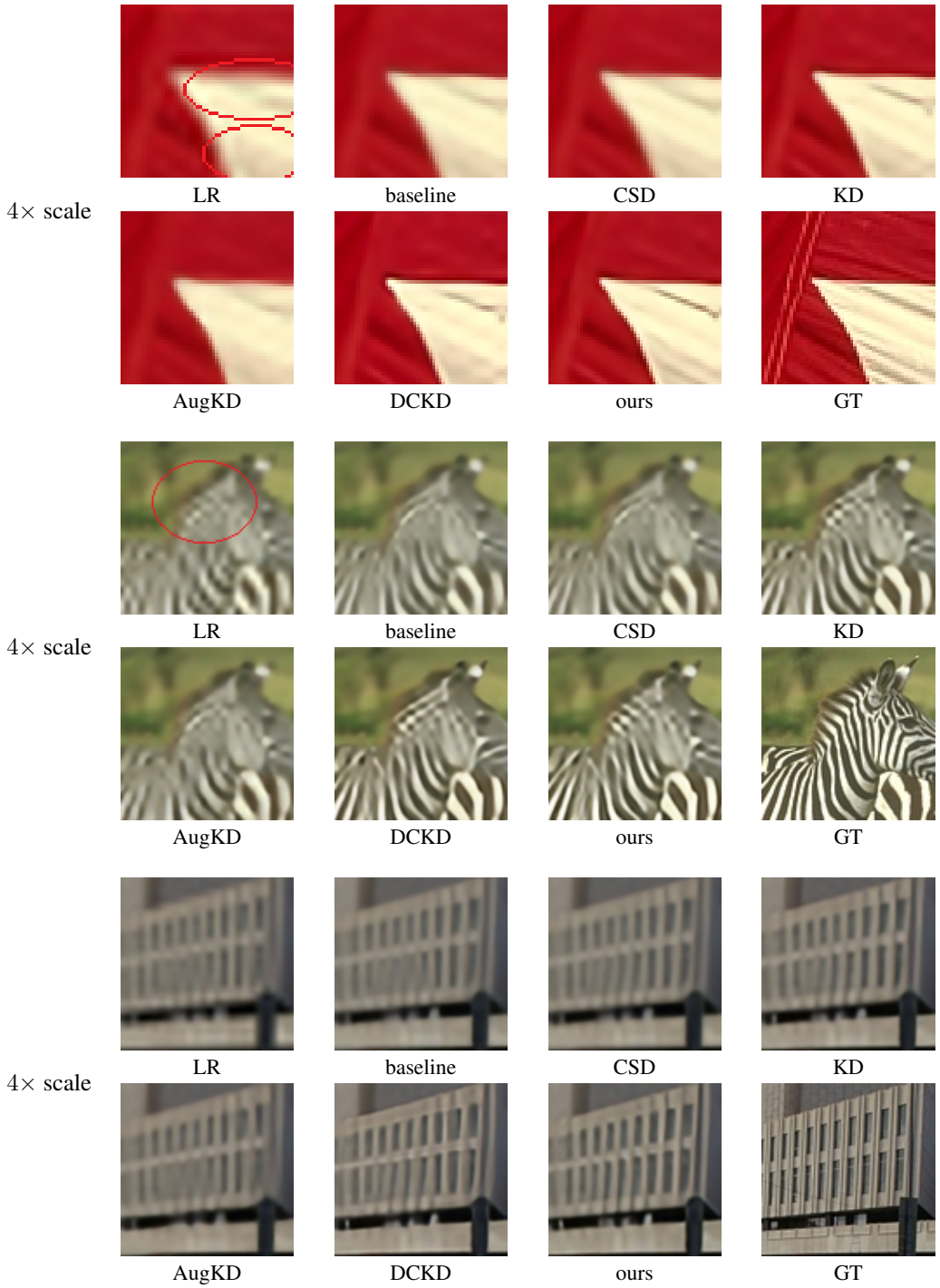


Figure 15: 4× qualitative comparison. **LR**: low-resolution input; **GT**: ground-truth. **baseline**: undistilled student. Columns labeled **CSD** (Wang et al., 2021b), **KD** (Hinton et al., 2015), **AugKD** (Zhang et al., 2025), and **DCKD** (Zhou et al., 2024a) show the student outputs distilled with the corresponding methods. **ours**: GAMD + ISCD. Fine-grained differences are best viewed when zoomed in.

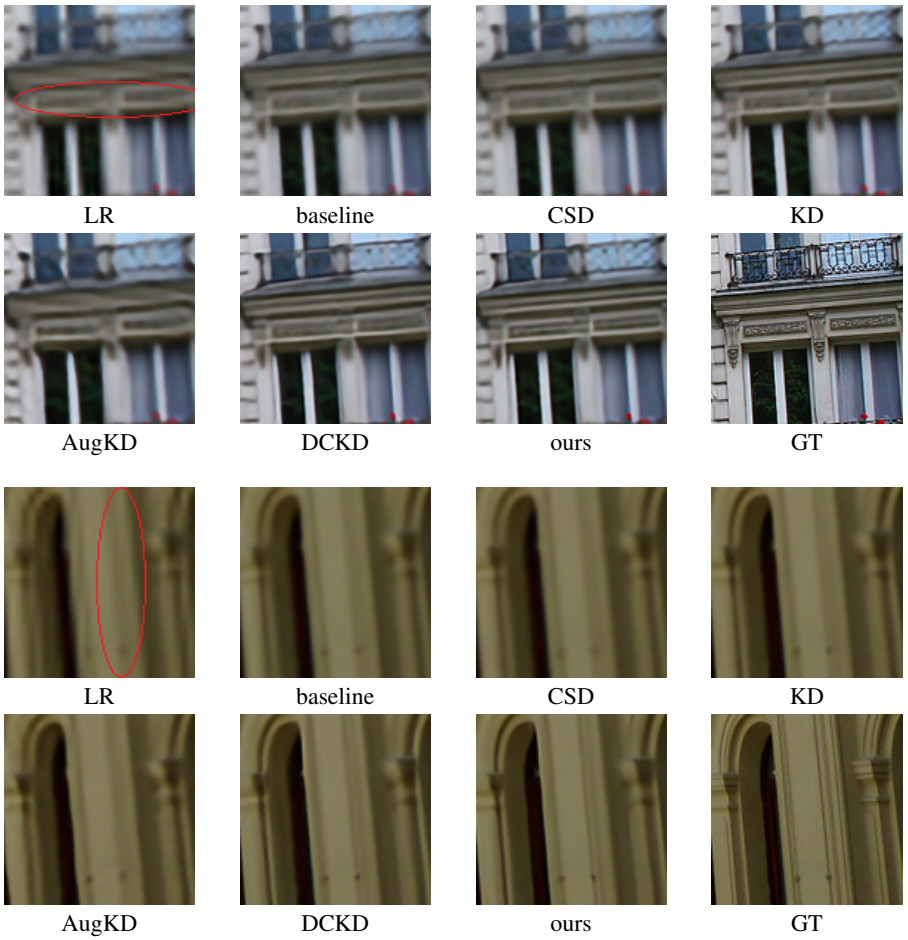


Figure 15: 4× qualitative comparison (continued).