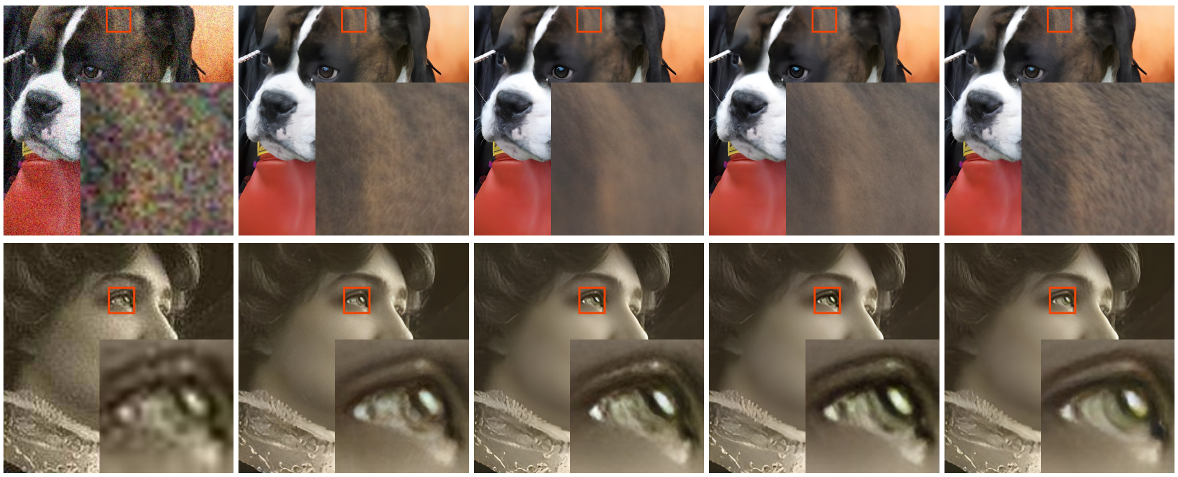
Auto-Encoded Supervision for Perceptual Image Super-Resolution

Supplementary Material



(a) Bicubic (b) BSRGAN (c) RealESRGAN (d) LDL (e) AESOP (Ours)

Figure 7. Visual comparison between AESOP (ours) against baseline methods for the real-world $\times 4$ SR task. AESOP leads to improved realism (top) with a lower level of visual artifacts (bottom). **Zoom in for best view.**

A. Implementation and experimental details

Network architecture and weight initialization. Following previous works, we initialize our SR networks with the official weight of the fidelity-oriented model of either ESRGAN [54] or SwinIR [36]. Similarly, the decoder of the AE follows the architecture of RRDB and is initialized with the fidelity-oriented weights. The overall architecture of the encoder is implemented in a straightforward manner. We simply design it as a series of two convolutional layers (fromRGB layer), followed by a pixel-unshuffle operation and two RRDB blocks [54], concluding with additional two convolutional layers (toRGB layer). The RRDB block is identical to that of the SR networks. The pixel-unshuffle acts as a $\times$s downscaling operation, effectively reducing the image dimension to match that of the LR image. Since the channel size is increased due to the pixel-shuffle operation, the second layer of the fromRGB layer reduces the channel size $\times s^2$ smaller than that of the RRDB block. The kernel size is 3$\times$3 for all convolutional layers.

Training and evaluation details. The optimizer is chosen as the Adam [26] optimizer with a learning rate of 0.0001, for both the Auto-Encoder and the SR network. Following conventions, we choose p = 1 for $\mathcal{L}_p$ and the coefficient of loss factors are $\lambda_{AESOP} = 1$, $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 0.005$. The Auto-Encoder is pretrained up to 100K iterations, and the SR networks are trained up to 300K iterations. Unless specified, the HR training patch size is 128. PSNR and SSIM scores are evaluated on the Y channel (luminance channel) in the YCbCr space and pixels up to the scale factors in the border were ignored. We use the default `alex` option for LPIPS [66]. Training and evaluation are performed on top of BasicSR [56]. Networks are trained and evaluated with either 4 NVIDIA A6000s or 4 NVIDIA RTX 3090s.

B. Evaluation on real-world SR datasets

AESOP on real-world SR. In the real-world SR task, the overall task becomes more complex and the range of plausible solutions is larger than that of the conventional bicubic SR task. Accordingly, the conflict between $\mathcal{L}_{pix}$ and perceptual quality-oriented objectives gets severe, and the blurring tendency of conventional $\mathcal{L}_{pix}$ loss may become more significant. We further validate the effectiveness of the proposed method in the real-world $\times 4$ SR task. For comparison, we use representative baseline real-world SR methods utilizing $\mathcal{L}_{pix}$, including RealESRGAN [55], BSRGAN [65], and LDL [37].

Qualitative results. In Fig.7, we provide a visual comparison of AESOP against baseline methods for the $\times 4$ real-world SISR task on RealSRSet [65]. We only replace the $\mathcal{L}_{pix}$ term of [37] while keeping all other training settings identical. Since we do not have ground-truth HR images, we only provide bicubic upsampled images and SR results from each method. Due to the inherent high complexity of the

Dataset	Method	NIQE↓	MANIQA	MUSIQ	CLIP-IQA
RealSRv3 [3]	ESRGAN [54]	7.7326	0.2043	29.0494	0.2362
	BSRGAN [65]	4.6519	0.3698	63.5908	0.5439
	Real-ESRGAN [55]	4.6790	0.3662	59.6855	0.4901
	LDL [37]	4.8869	0.3706	60.1015	0.4883
	AESOP (Ours)	**4.2337**	**0.4136**	**63.6489**	**0.5687**
DRealSR [58]	ESRGAN [54]	8.3949	0.2115	20.2861	0.2468
	BSRGAN [65]	4.6809	0.3381	35.4973	0.5614
	Real-ESRGAN [55]	4.7152	0.3404	35.2747	0.5098
	LDL [37]	5.0974	0.3393	35.9026	0.5137
	AESOP (Ours)	**4.1922**	**0.3917**	**36.5533**	**0.6025**

Table 6. Quantitative results of AESOP and baseline methods in real-world settings. All methods except ESRGAN are trained for the real-world SR task. The best results of each group are highlighted in **bold**. ↓ means lower is better. If not specified, higher is better. Due to memory constraints, images were cropped before evaluating CLIP-IQA scores for the DRealSR dataset.

real-world task, baseline networks fail in generating fine-grained textures (first row of Fig.7) and generate visually unpleasing artifacts (second row of Fig.7). In contrast, AESOP successfully recovers fine textures with fewer artifacts.

Quantitative results. We report quantitative results on RealSRv3 [3] and DRealSR [58]. To assess perceptual quality, we utilize NIQE [46], MANIQA [60], MUSIQ [24], and CLIP-IQA [52] scores. Due to memory constraints, images were divided into four quadrants when evaluating the CLIP-IQA scores for the DRealSR dataset. AESOP demonstrates superior performance against baselines in all evaluation metrics, which verifies the effectiveness of our method for practical applications.

C. Additional results for the Bicubic SR task

FID scores. In Tab.7, we report Frechet Inception Distance (FID) [17] scores to further evaluate the proposed AESOP against baseline methods for the bicubic $\times 4$ SR task. FID, widely used for generative tasks [23], has recently been adopted for super-resolution tasks [37, 48]. However, its standard approach of resizing images to 299×299 may not be suitable to assess SR methods. Resizing can alter important details that SR aims to improve, directly conflicting with the objectives of SR focusing on enhancing image quality at higher resolutions.

Patch FID scores. Accordingly, we additionally report the patch-FID (pFID) [4] scores, which does not require image resizing. For patch-FID evaluation, 299×299 non-overlapping patches are extracted from the images. If an image is smaller than 299 pixels in any dimension, we use zero-padding to meet the required size.

Fidelity bias estimation. As discussed in the main article, we do not multiply a small scaling factor to $\mathcal{L}_{\text{AESOP}}$ which leads to significantly stronger guidance on fidelity biases (Fig.9). Accordingly, we have measured how well AESOP and the baseline methods estimate the fidelity biases by reporting AE-PSNR which captures the distance between the fidelity bias of the SR image and the fidelity bias of the HR image. Additionally, we have shown LR-PSNR scores to provide a metric that is not biased by the Auto-Encoder. In Tab.8, we additionally provide AE-PSNR and LR-PSNR scores on top of the RRDB [54] backbone. Similar to results in Tab.4, AESOP shows improvements in both AE-PSNR and LR-PSNR scores, highlighting the superiority of AESOP in effectively reducing the SE term.

AESOP on recent backbone network architecture. We report additional quantitative results on the benchmark datasets in Tab.9. First, we employ DRCT [18], a recent state-of-the-art Swin Transformer-based method that leverages dense residual connections within a fidelity-oriented SR framework. We implement LDL on top of DRCT and compare it to our proposed AESOP. AESOP consistently outperforms the baseline in terms of both fidelity and perceptual quality, demonstrating its effectiveness even with advanced network architectures. Notably, the performance improvement is more significant compared to the RRDB backbone, suggesting that AESOP may yield even greater benefits with larger-capacity network architectures.

Regarding recent perceptual-oriented losses. We report quantitative results of another recent state-of-the-art method, CALGAN [48]. This work is a different branches of research in the field of perceptual SR, focusing on improvements in perceptual quality-oriented losses. Interestingly, AESOP outperforms CALGAN in most cases, even without the Mixture of Experts (MoE)–based discriminator proposed in CALGAN [48]. This signifies the effectiveness of AESOP. However, note that improvements in network architectures and perceptual-oriented losses are beyond the scope of this work. The focus of this study is on the fidelity loss term $\mathcal{L}_{\text{pix}}$ within the perceptual SR framework. We leave the integration of $\mathcal{L}_{\text{AESOP}}$ (fundamentally a *fidelity* loss), with the enhanced perceptual-oriented losses of CALGAN to future work due to limited computational budget.

Metrics	Backbone	RRDB						SwinIR			
Metrics	Benchmark	ESRGAN	SPSR	LDL*	LDL	AESOP	AESOP	+GAN	LDL*	LDL	AESOP
Recon. Objective		$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\textbf{AESOP}}$	$\mathcal{L}_{\textbf{AESOP}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\textbf{AESOP}}$
Patch size (Training)		128	128	128	128	128	256	256	256	256	256
pFID $\downarrow$	Set14	65.220	70.990	-	57.132	**56.727**	54.792	-	-	55.367	**53.175**
	Manga109	29.326	28.314	-	23.895	**23.384**	22.833	-	-	21.766	**21.290**
	General100	50.062	50.053	-	43.406	**42.117**	41.041	-	-	42.028	**40.199**
	Urban100	32.094	31.105	-	28.380	**27.875**	27.017	-	-	26.972	**25.613**
	BSD100	69.943	68.370	-	64.058	**57.864**	56.844	-	-	59.653	**57.118**
	LSDIR	14.579	14.110	-	12.537	**12.220**	11.718	-	-	12.056	**11.387**
FID $\downarrow$	Set14	54.939	53.919	43.454	**43.479**	46.828	38.907	48.910	46.057	46.110	**45.411**
	Manga109	11.559	10.663	10.161	10.162	**9.230**	9.446	9.703	8.680	**8.677**	9.256
	General100	29.850	30.172	27.211	**27.220**	27.425	25.201	27.557	25.304	25.301	**24.592**
	Urban100	20.354	18.676	16.351	**16.355**	16.865	16.723	17.555	16.282	16.287	**15.547**
	BSD100	50.752	48.349	-	44.053	**41.130**	40.751	-	-	41.954	**41.721**
	LSDIR	17.552	16.056	-	15.229	**14.748**	14.802	-	-	14.510	**14.397**

Table 7. Quantitative results of the proposed method (AESOP) against baseline methods. We report Frechet Inception Distance (FID) and patch-FID (pFID) scores. LDL* indicates that scores are from the official paper. All other scores are evaluated in our settings, with officially provided pretrained weights. The best results of each group are highlighted in **bold**, based on scores evaluated in our settings.

Metric	Method	Set14	Manga109	General100	Urban100	DIV2K-val	BSD100	LSDIR
AE-PSNR	ESRGAN [54]	30.280	31.165	32.663	27.198	31.668	28.991	27.636
	SPSR [42]	30.602	31.351	32.670	27.508	31.737	29.029	27.881
	LDL [37]	31.180	32.608	33.823	28.488	32.597	29.595	28.625
	AESOP (Ours)	**31.341**	**32.843**	**33.956**	**28.529**	**32.740**	**29.737**	**28.812**
LR-PSNR	ESRGAN [54]	43.892	43.908	45.259	42.879	45.689	43.823	42.718
	SPSR [42]	43.835	44.359	44.656	42.666	44.717	42.719	42.364
	LDL [37]	46.497	47.603	48.184	45.975	47.793	45.307	45.295
	AESOP (Ours)	**46.625**	**48.188**	**48.653**	**46.280**	**48.272**	**45.837**	**45.571**

Table 8. Quantitative comparison between the proposed method (AESOP) and baseline methods. We report AE-PSNR and LR-PSNR scores using the RRDB backbone. AE-PSNR measures how accurately the method estimates fidelity bias factors, while LR-PSNR evaluates how well the generated images align with the input LR image. The best result in each group is highlighted in bold.

D. Further discussion on AESOP

Regarding the loss maps and spectral magnitudes. Here we provide further discussions regarding the loss maps and the spectral analysis in the main article. In Sec.5.2, we have discussed the differences between AESOP and *low*-pass filtering-based methods. However, the loss maps reveal object edges, which are the regressable high-frequency components, aligning to *high*-pass filters. Accordingly, we provide further discussion and compare AESOP against high-pass filter based losses or similarly against edge filters from two perspectives: 1) regions with low loss values under $\mathcal{L}_{\text{AESOP}}$ and 2) regions with high loss values under HPF losses. (Fig.8)

First, we emphasize that regions with low loss values under $\mathcal{L}_{\text{AESOP}}$ do not imply that $\mathcal{L}_{\text{AESOP}}$ neglects these areas. Instead, they simply indicate that the network has accurately estimated the fidelity bias in those regions. This is clearly different from frequency filters, which entirely ignore these regions. For instance, consider a scenario where the SR network produces low-frequency artifacts due to adversarial training instability. In such cases, $\mathcal{L}_{\text{AESOP}}$ effectively guides the network toward proper estimation, whereas HPF loss ignores these artifacts, resulting in suboptimal performance. This also suggests that the components that require reconstruction guidance and those that do not require reconstruction guidance are inherently intertwined within each pixel. Thus, they cannot be disentangled merely by selecting which pixels to penalize.

Meanwhile, for regions that receive high loss activations under high-pass filtering (HPF) loss, these typically correspond to areas with fine textures. This is exactly the problematic issue raised in $\mathcal{L}_{\text{pix}}$, where such activations con-

Backbone		RRDB		SwinIR		DRCT	
Metrics	Benchmark	CALGAN	AESOP	CALGAN	AESOP	LDL	AESOP
Recon. Objective		$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{AESOP}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{AESOP}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{AESOP}}$
LPIPS ↓	Set14	-	0.1053	-	0.1027	0.1086	**0.1022**
	Manga109	-	0.0494	-	0.0461	0.0459	**0.0447**
	General100	0.077	**0.0734**	0.074	**0.0710**	0.0727	**0.0722**
	Urban100	0.108	**0.1033**	0.098	**0.0945**	0.1006	**0.0972**
	DIV2K-val	**0.091**	0.0936	**0.087**	0.0893	**0.0934**	0.0949
	BSD100	0.151	**0.1443**	0.147	**0.1385**	0.1462	**0.1451**
	LSDIR	-	0.1123	-	0.1071	0.1131	**0.1129**
DISTS ↓	Set14	-	0.0825	-	0.0819	0.0889	**0.0830**
	Manga109	-	0.0356	-	0.0328	**0.0316**	0.0338
	General100	0.083	**0.0773**	0.081	**0.0762**	0.0782	**0.0775**
	Urban100	0.082	**0.0768**	0.083	**0.0742**	0.0803	**0.0771**
	DIV2K-val	0.049	**0.0484**	0.048	**0.0459**	0.0487	**0.0485**
	BSD100	0.118	**0.1089**	0.128	**0.1072**	0.1136	**0.1072**
	LSDIR	-	0.0612	-	0.0591	0.0635	**0.0621**
PSNR ↑	Set14	-	27.246	-	27.421	27.314	**27.796**
	Manga109	-	29.747	-	30.061	29.979	**30.398**
	General100	30.182	**30.251**	-	30.401	30.143	**30.646**
	Urban100	25.290	**25.541**	-	26.148	26.038	**26.360**
	DIV2K-val	28.863	**28.910**	-	29.137	29.030	**29.456**
	BSD100	**25.925**	25.904	-	25.930	25.942	**26.324**
	LSDIR	-	24.845	-	25.038	24.943	**25.354**
SSIM ↑	Set14	-	0.7371	-	0.7438	0.7403	**0.7546**
	Manga109	-	0.8802	-	0.8880	0.8888	**0.8936**
	General100	0.825	**0.8269**	-	0.8327	0.8288	**0.8382**
	Urban100	0.763	**0.7697**	-	0.7884	0.7855	**0.7926**
	DIV2K-val	0.790	**0.7951**	-	0.8023	0.7994	**0.8085**
	BSD100	0.676	**0.6783**	-	0.6813	0.6812	**0.6921**
	LSDIR	-	0.7202	-	0.7289	0.7253	**0.7353**

Table 9. Additional quantitative evaluation on benchmark datasets. We also provide quantitative results of CALGAN [48] and DRCT [18]. CALGAN is a recent work improving perceptual-oriented losses, while DRCT made improvements in the SR network architecture. AESOP mostly outperforms CALGAN [48] even without the MoE-discriminator. However, note that enhancements to network architectures and perceptual-oriented losses are beyond the scope of this work. The focus of this work is on the fidelity loss term $\mathcal{L}_{\text{pix}}$ under the perceptual SR framework. The best results of each group are highlighted in **bold**. Additionally, refer to the PD trade-off curve in Fig.14-20.

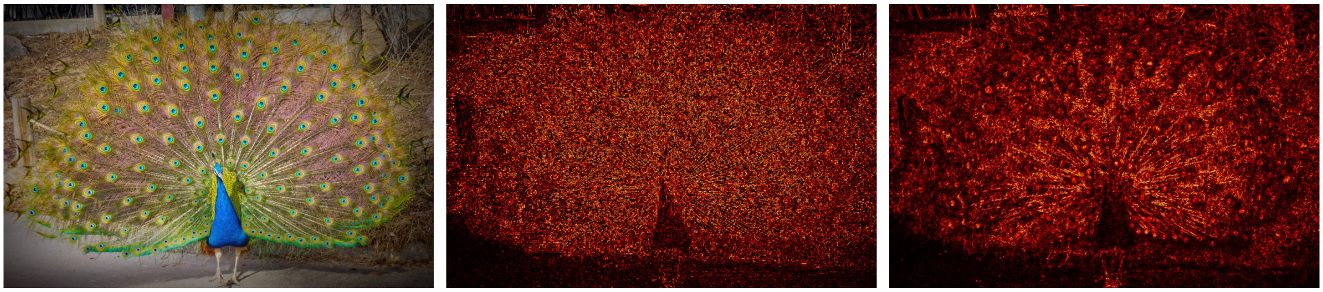
tribute to blurring. Consequently, this represents an undesirable aspect of HPF-based methods.

Intuitions on $\mathcal{L}_{\text{AESOP}}$ based on loss scales. In Fig.9, we compare the loss scales of $\mathcal{L}_{\text{pix}}$ and $\mathcal{L}_{\text{AESOP}}$, both before and after applying their loss coefficients. Before the loss coefficients are applied, $\mathcal{L}_{\text{pix}}$ (green) exhibits greater loss values than $\mathcal{L}_{\text{AESOP}}$ (blue). This observation aligns with our theoretical analysis and construction of the Auto-Encoder, where $\mathcal{L}_{\text{AESOP}}$ only penalizes a subcomponent of $\mathcal{L}_{\text{pix}}$. Specifically, while $\mathcal{L}_{\text{pix}}$ minimizes both perceptual variance (VE) and fidelity bias induced error (SE), our carefully designed $\mathcal{L}_{\text{AESOP}}$ only targets the SE term, leading to lower loss values. Consequently, the gap between the green loss trajectory and the blue one quantifies the VE loss component embedded within $\mathcal{L}_{\text{pix}}$. After the loss coefficients are applied to each reconstruction loss, $\mathcal{L}_{\text{AESOP}}$ (blue) provides an order of magnitude stronger reconstruction guidance compared to scaled $\mathcal{L}_{\text{pix}}$ (red). Regardless of this strengthened fidelity guidance, SR networks trained with $\mathcal{L}_{\text{AESOP}}$ do not have to suffer from blurring and can achieve improved perceptual quality over $\mathcal{L}_{\text{pix}}$.

Intuitions on $\mathcal{L}_{\text{pix}}$ and $\mathcal{L}_{\text{AESOP}}$. Apart from Fig.1, we show additional graphical illustration in Fig.11 to provide further intuitions on the overall optimization procedure and the optimal point of each $\mathcal{L}_{\text{pix}}$ and $\mathcal{L}_{\text{AESOP}}$. As can be seen, $\mathcal{L}_{\text{AESOP}}$ consecutively estimates the centroid (fidelity bias) of the prediction and solution space, and minimizes the distance between them (i.e., minimizes the SE factor). Accordingly, $\mathcal{L}_{\text{AESOP}}$ reaches the optimal point when the two distributions are aligned. However, $\mathcal{L}_{\text{pix}}$ converges to the minimum expected error point, which is the blurry average solution. Thus, the prediction space degenerates.

Comparison between $\mathcal{L}_{\text{percep}}$ and $\mathcal{L}_{\text{AESOP}}$. The proposed loss $\mathcal{L}_{\text{AESOP}}$ and the perceptual loss $\mathcal{L}_{\text{percep}}$ share the characteristic of utilizing a pretrained neural network for guidance. However, they differ fundamentally in their objectives and mechanisms. Below, we clarify these differences in two different aspects.

First, the primary objectives of these two losses differ significantly.$\mathcal{L}_{\text{percep}}$ belongs to the category of perceptual-oriented losses. Its main purpose is to explicitly improve perceptual quality by providing *high-level* semantics and



(a) Original (b) High-pass filtered loss (c) L_{AESOP}

Figure 8. Loss map comparison between high-pass filtered (HPF) loss and $\mathcal{L}_{\text{AESOP}}$. Refer to Appendix.D for further discussion.

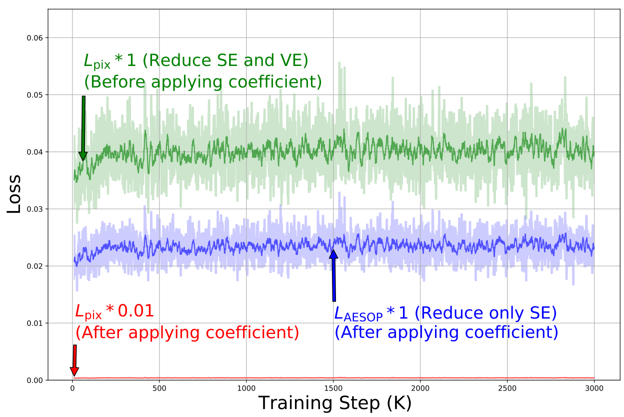


Figure 9. The loss trajectory of $\mathcal{L}_{\text{pix}}$ before applying the coefficient (green) is visualized by scaling the original loss (red). The loss trajectory of $\mathcal{L}_{\text{AESOP}}$ (blue) is visualized as-is, since we do not scale it. Refer to Appendix.D for further discussion.

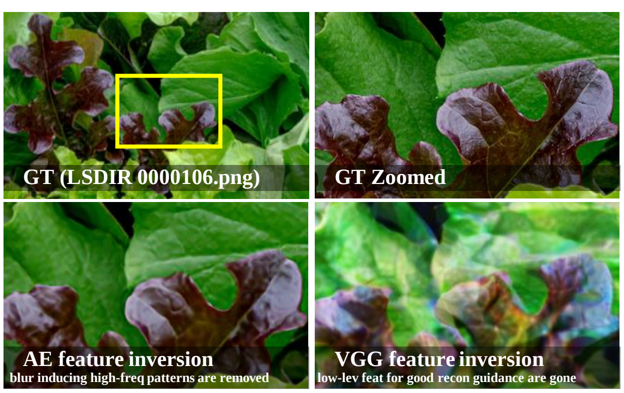
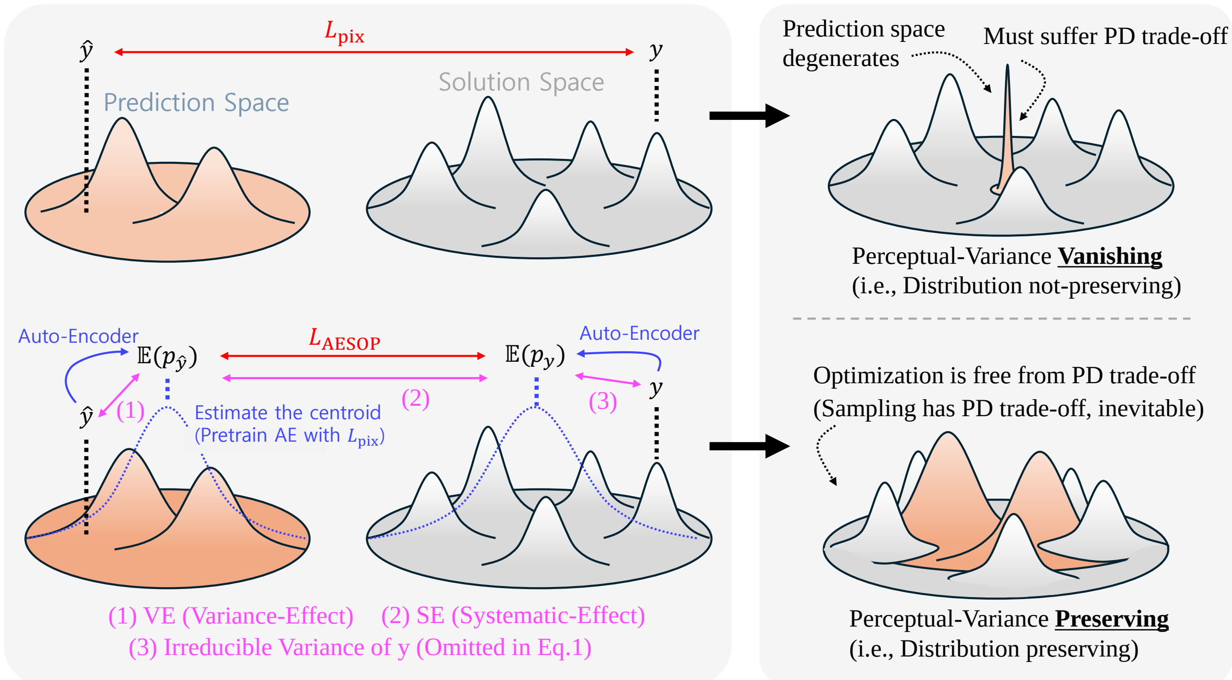


Figure 10. Comparison between feature inversion results obtained from deep features extracted by VGG and our proposed AE. Deep features of VGG lose important low-level features crucial for a reconstruction loss. Meanwhile, the carefully chosen network architecture and the pretraining objective of our AE enable precise control over which information to remove (blur-inducing high-frequency patterns) and preserve (structural edges).

textural guidance. Accordingly, $\mathcal{L}_{\text{percep}}$ measures the discrepancy between the SR and HR images within a high-dimensional feature space derived from a pretrained feature extractor (such as VGG [51]), where the high-dimensional space captures additional semantic and textural details beyond those available in the raw pixel domain. In contrast, $\mathcal{L}_{\text{AESOP}}$ is fundamentally a reconstruction (fidelity) loss that provides guidance based on low-level features, similar to the conventional $\mathcal{L}_{\text{pix}}$, but specifically tailored for perceptual SR tasks so that it does not show conflicts with perceptual-oriented losses. $\mathcal{L}_{\text{AESOP}}$ employs an Auto-Encoder (AE) architecture with a low-dimensional bottleneck, pretrained for low-level reconstruction. Due to its design and pretraining objective, the AE inherently compresses the input and selectively discards certain information, while preserving important low-level features. Consequently, the Auto-Encoded output contains *less* information compared to the original image, as opposed to the enriched, high-dimensional features used in $\mathcal{L}_{\text{percep}}$.

Second, the underlying mechanism and the information each feature encoder embeds are different. In order to utilize a feature encoder as a loss function in low-level vision tasks, precise control over which information to remove and preserve is important. Considering that a reconstruction loss in perceptual SR task should (1) provide sufficient reconstruction guidance while (2) avoid blurring; feature encoders should be able to preserve important low-level features while removing blur-inducing factors. However, feature encoders pretrained on image classification tasks (such as VGG) naturally discard many low-level features not relevant to classification, resulting in uncontrollable loss of critical reconstruction information. In contrast, the carefully designed AE preserves essential low-level features, particularly structural edges, while the blur inducing perceptual variance factors are removed.

We empirically verify these properties through feature inversion results shown in Fig.10. Clearly, deep features extracted from VGG omit critical low-level reconstruction details. On the other hand, our AE-derived deep features successfully retain sharp edges and structural alignment while



(a) Comparison between L_{pix} and L_{AESOP}

(b) Optimal point of L_{pix} and L_{AESOP}

Figure 11. Graphical illustration of the optimization procedure and the optimal point for $\mathcal{L}_{\text{pix}}$ and $\mathcal{L}_{\text{AESOP}}$.

the blur inducing high-frequency textural information is removed as intended. Overall, this verifies that $\mathcal{L}_{\text{percep}}$ cannot act as a standalone reconstruction loss in low-level vision tasks, while $\mathcal{L}_{\text{AESOP}}$ can, and is even shown to outperform conventional $\mathcal{L}_{\text{pix}}$ through extensive experiments.

Disclaimer. We clarify that the improvement in perceptual scores by raising $\mathcal{L}_{\text{AESOP}}$ is since it does not hinder the perceptual-oriented guidance provided by perceptual-oriented losses under the SRGAN-framework. $\mathcal{L}_{\text{AESOP}}$ itself will not guide towards realism. We keep improvements in perceptual-oriented losses out of the scope of this work.

E. Further intuition regarding the PD trade-off

Comparison between $\mathcal{L}_{\text{pix}}$ and $\mathcal{L}_{\text{AESOP}}$. Fig.12 represents the guidance $\mathcal{L}_{\text{pix}}$ and $\mathcal{L}_{\text{AESOP}}$ provides in terms the perception-distortion (PD) trade-off. We start our discussion with point (B), which represents an image that is not optimal in both fidelity and perception. Given this image, $\mathcal{L}_{\text{pix}}$ with a large coefficient guides the image towards point (C). This is the blurry image with the lowest expected distortion, or simply the fidelity bias of the image. Meanwhile, with a smaller coefficient, it achieves improved perception as point (G). However, it leads to unnecessary fidelity loss (H) since SE reduction is significantly weakened while the adversarial loss continuously hinders SE convergence. Meanwhile, $\mathcal{L}_{\text{AESOP}}$ removes the VE minimization term of $\mathcal{L}_{\text{pix}}$. Thus, it improves fidelity without suffering from blurring, thereby guides point (B) towards point (A). However, we clarify that $\mathcal{L}_{\text{AESOP}}$ cannot further improve the fidelity beyond the ideal PD trade-off curve. This is impossible as (E), under non-invertible degradation [2] including image super-resolution. This statement even holds for the case with an optimal perceptual SR network that can sample images from the true posterior. Note that $\mathcal{L}_{\text{AESOP}}$ reaches zero for point (A).

Is $\mathcal{L}_{\text{AESOP}}$ a distortion measure? Blau et. al. [2] have shown that we must compensate perception when aiming to reduce *any* distortion measure; the perception-distortion trade-off. This might seem contradictory with $\mathcal{L}_{\text{AESOP}}$ at first glance, since $\mathcal{L}_{\text{AESOP}}$ is designed to improve fidelity without degrading perception. However, fortunately, $\mathcal{L}_{\text{AESOP}}$ does not fall within the definition of distortion metric defined in Blau et. al. [2]. A distortion measure Δ that induces PD trade-off requires: $\Delta(y_1, y_2) > 0$ for $y_1 \neq y_2$ by definition. However, for AESOP, it is straightforward (and also intended) that multiple different images can share an identical fidelity bias. Formally, there exists y_1, y_2 s.t. $\mathcal{L}_{\text{AESOP}}(y_1, y_2) = 0$ and $y_1 \neq y_2$. As this does not satisfy the constraints of a distortion measure, $\mathcal{L}_{\text{AESOP}}$ is not guaranteed to raise PD trade-off. However, we clarify that this does not imply that SR networks trained with $\mathcal{L}_{\text{AESOP}}$ can

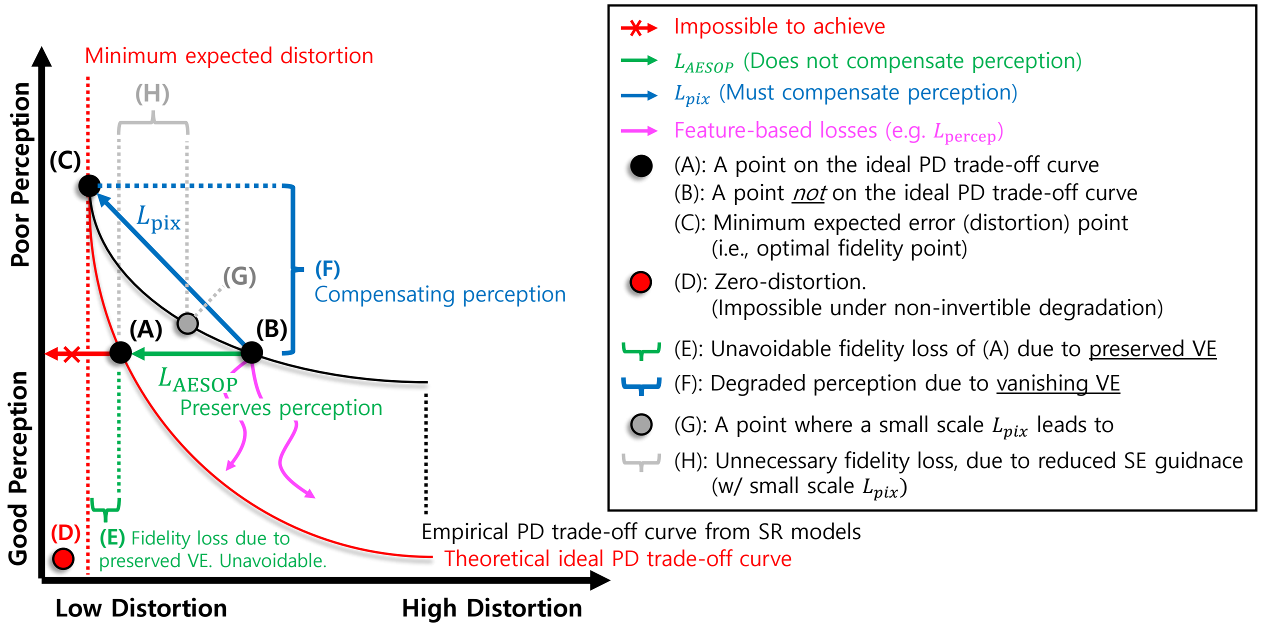


Figure 12. Graphical illustration of $\mathcal{L}_{\text{pix}}$ and $\mathcal{L}_{\text{AESOP}}$ in terms of the perception-distortion trade-off.

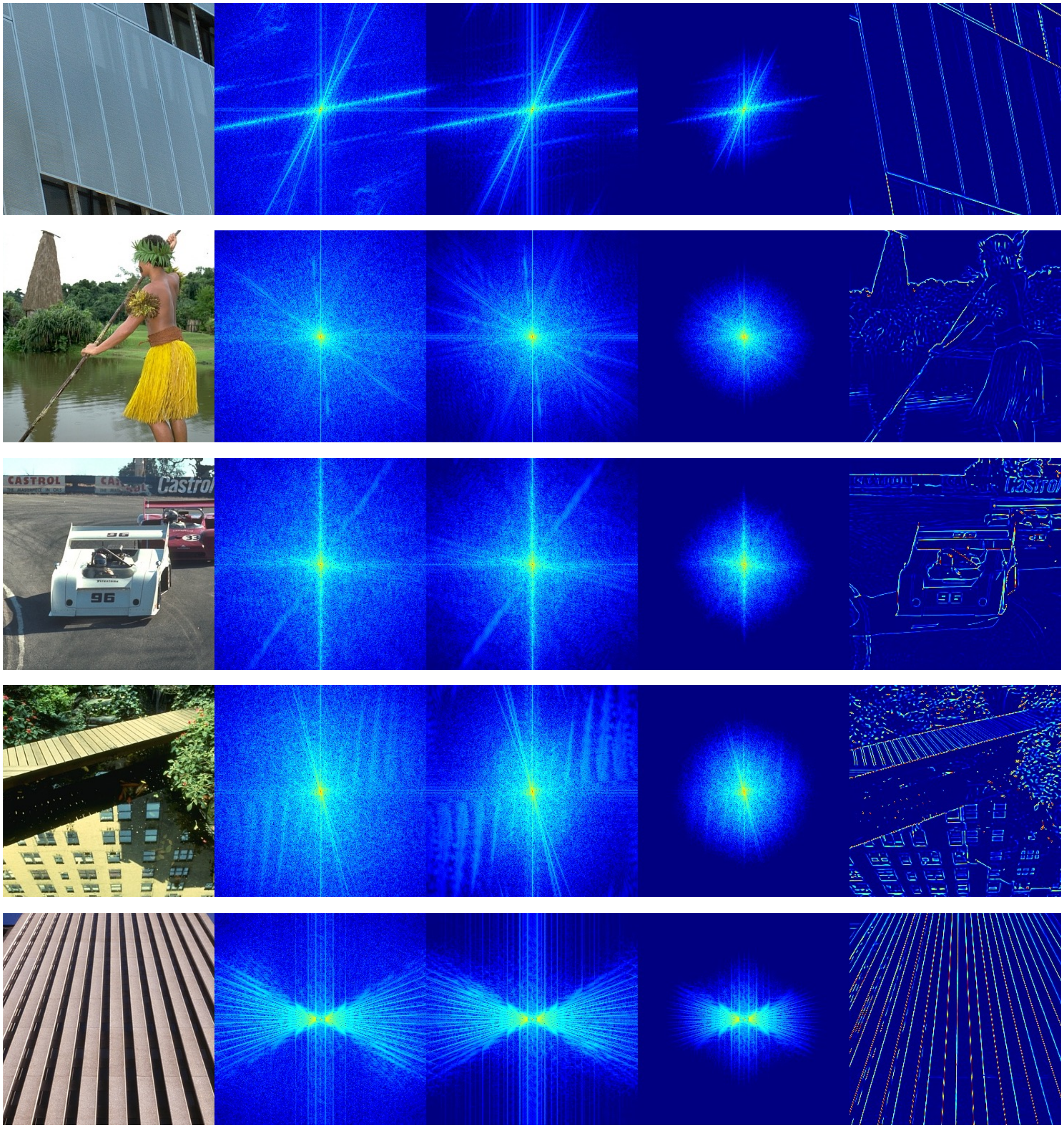
generate images that are free from the PD trade-off. This is impossible even with an oracle perceptual SR network, as discussed in prior sections.

F. Additional visualization

Spectral magnitudes. Fig.13 provides visual examples of the spectral magnitudes, aligning with Fig.6. The spectral magnitudes reflect the effectiveness of the pretrained Auto-Encoder in discriminating non-regressable factors that lead to blurring and the regressable high-frequency components that enhance fidelity without causing blurring. Meanwhile, low-pass filters fail to achieve this since the regressable and non-regressable components cannot be disentangled using simple frequency filters. They are intertwined within the same frequency band.

Qualitative examples on benchmark datasets. To further illustrate the effectiveness of our method, we present an additional qualitative comparison between AESOP against the baseline method LDL [37] on the bicubic $\times 4$ SR task. We provide results of tested methods, AESOP and LDL, on top of the SwinIR backbone (Fig.21 and Fig.22) and the RRDB backbone (Fig.23 and Fig.24). As can be seen, AESOP significantly improves perceptual quality while effectively suppressing visual artifacts observed in the baseline method.

Additional perception-distortion trade-off curves. We provide extensive visualizations of the perception-distortion trade-off curves in Fig.14-20. For CALGAN [48], we present only a single data point rather than the full perception-distortion trade-off curve, as its official weights are not publicly available. Extensive results show that AESOP leads to substantial performance improvements against baselines in terms of the perception-distortion trade-off. Aligning to Tab.9, AESOP also often outperforms CALGAN even without MoE-discriminator proposed in CALGAN. Additionally, we observe that AESOP often results in larger improvements for Swin Transformer-based methods (e.g., SwinIR, DRCT) compared to CNN-based methods (e.g., RRDB). This is likely because these models have greater capacity and benefit more from the enhanced reconstruction guidance provided by AESOP. However, there are instances where AESOP does not always lead to improved performance. Specifically, AESOP often fails to enhance performance on the Manga109 [44] dataset, which is consistent with the unexpected trade-off behaviors observed across most methods in this dataset. This limitation arises because Manga109 consists predominantly of comic images, which typically lack the fine-grained textures found in photorealistic datasets. The absence of such textures poses a challenge for perceptual SR methods, including AESOP, which are specifically designed to enhance and preserve realistic textures. Consequently, without the presence of these detailed textures, AESOP's advantages in minimizing fidelity bias and preserving perceptual variance are less pronounced, leading to suboptimal performance in this particular dataset.



(a) Original (b) Original (c) AE (d) LPF (e) Difference

Figure 13. Visual comparison between Auto-Encoding and low-pass filtering. **(a)** Original image. **(b)** Original image in spectral domain. **(c)** Auto-Encoded image. **(d)** Low-pass filtered image. **(e)** Absolute difference between the Auto-Encoded image and the low-pass filtered image. **Electronic viewer recommended**.

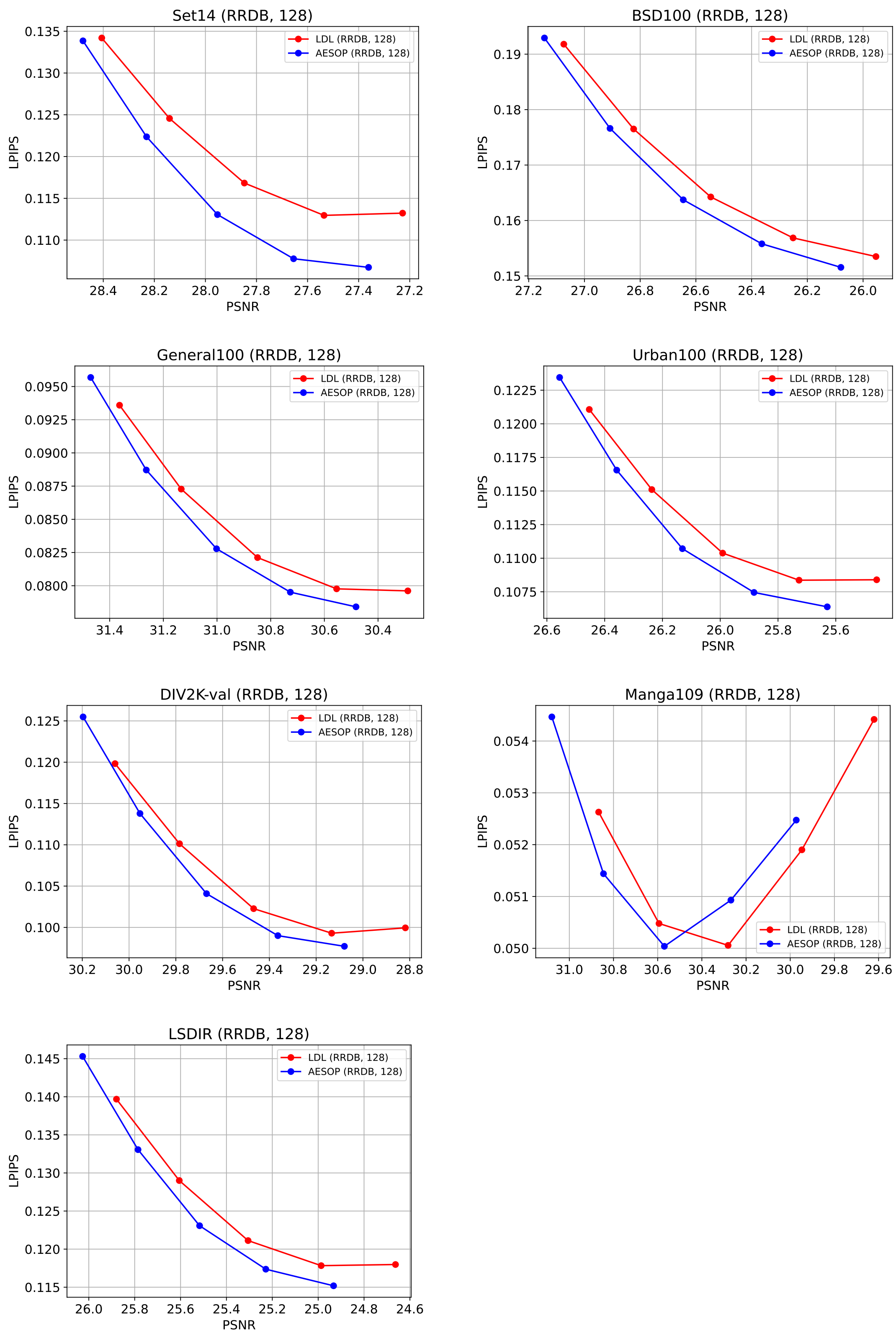


Figure 14. The perception-distortion trade-off curve between AESOP and baseline methods on top of the RRDB [54] backbone. The training HR patch size is 128. AESOP often fails to improve the performance on the Manga109 dataset.

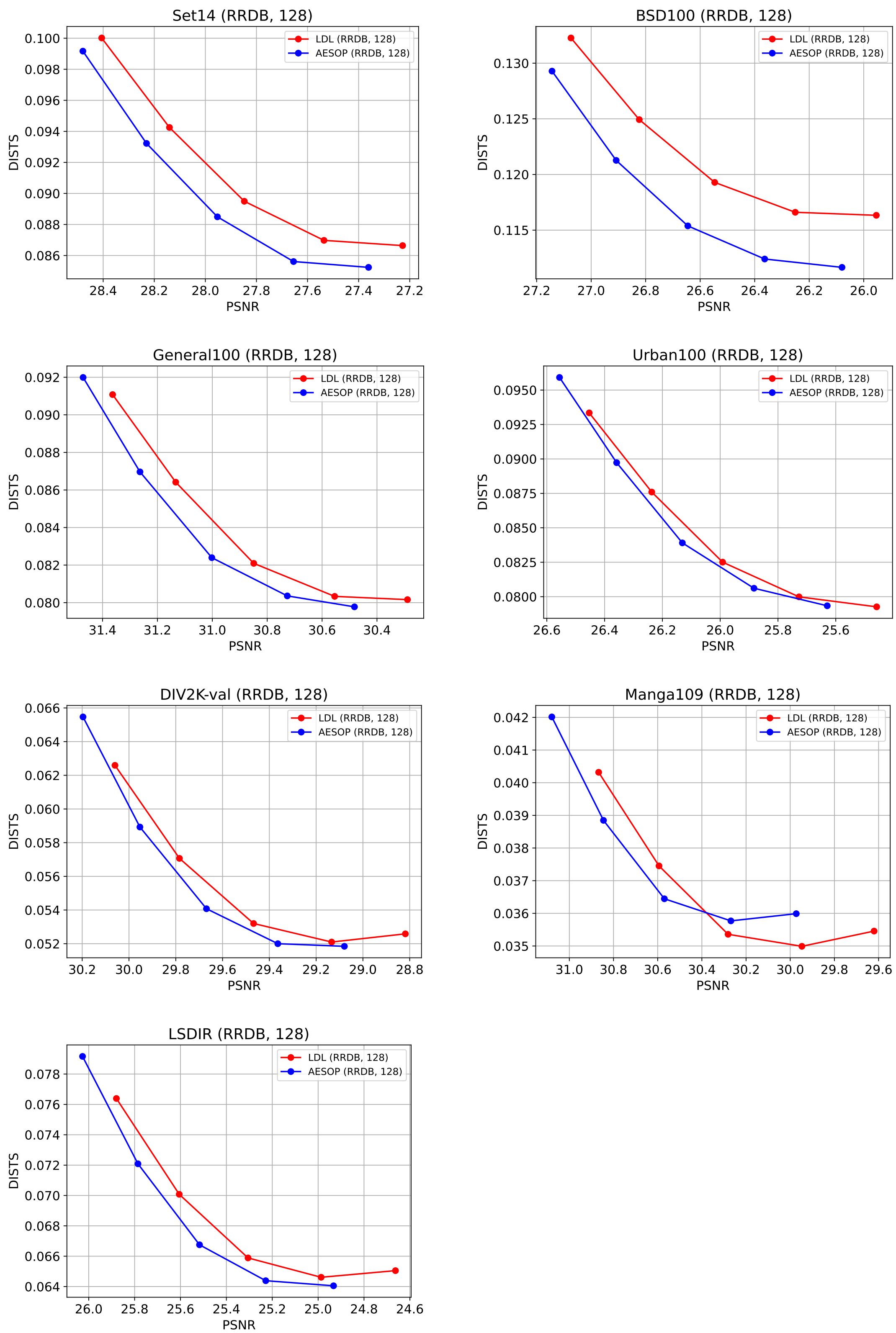


Figure 15. The perception-distortion trade-off curve between AESOP and baseline methods on top of the RRDB [54] backbone. The training HR patch size is 128. AESOP often fails to improve the performance on the Manga109 dataset.

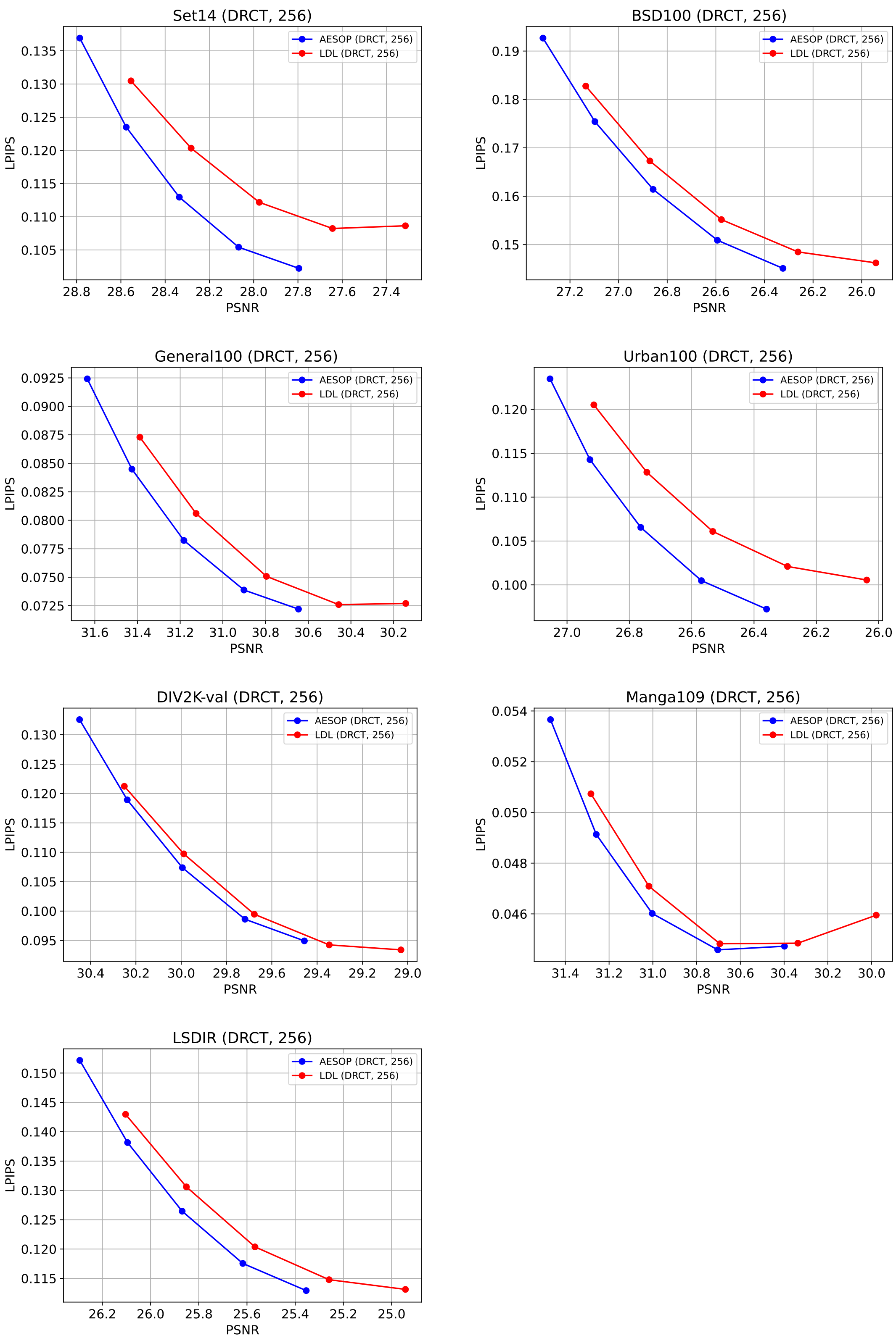


Figure 16. The perception-distortion trade-off curve between AESOP and baseline methods on top of the DRCT [18] backbone. The training HR patch size is 256. AESOP often fails to improve the performance on the Manga109 dataset.

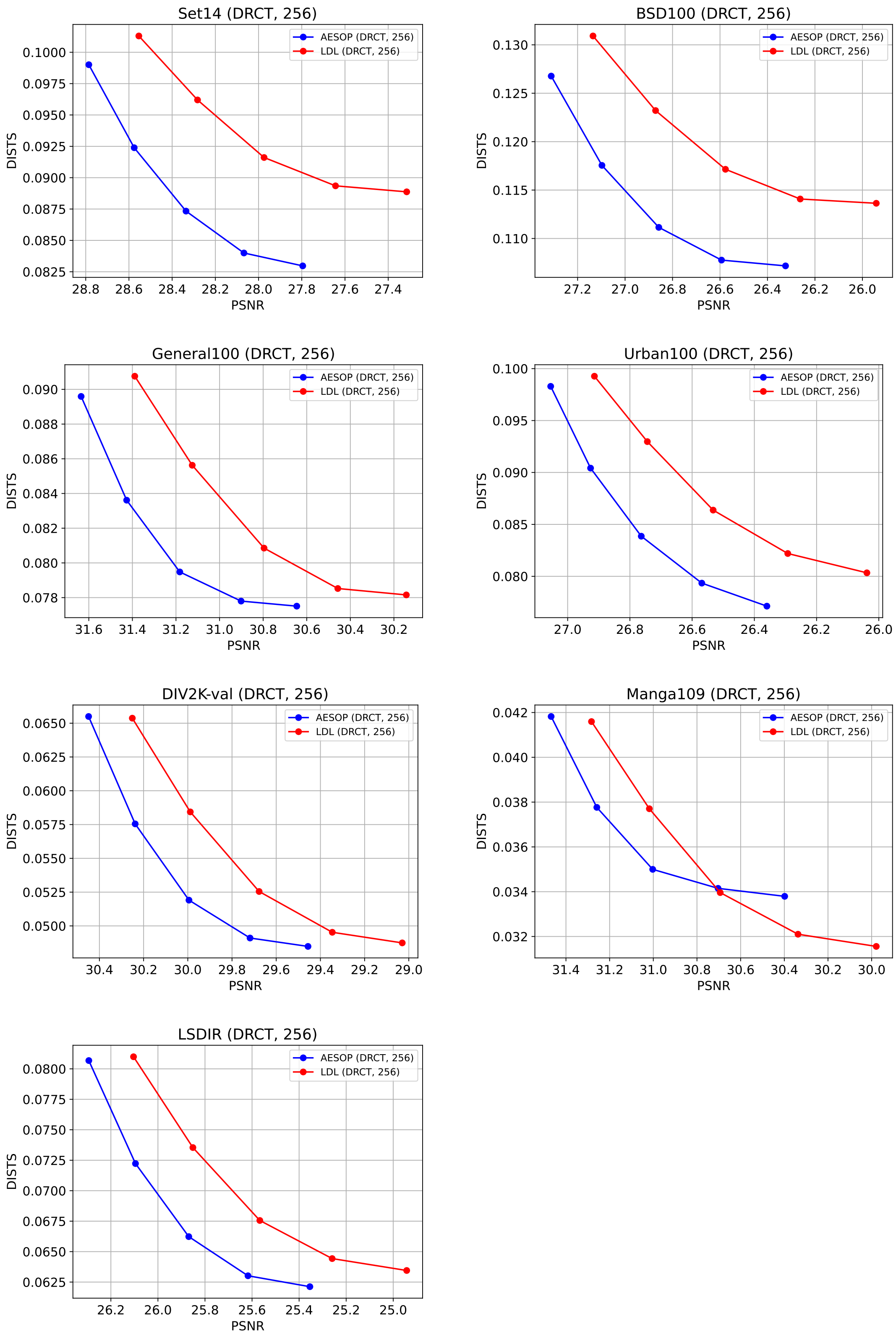


Figure 17. The perception-distortion trade-off curve between AESOP and baseline methods on top of the DRCT [18] backbone. The training HR patch size is 256. AESOP often fails to improve the performance on the Manga109 dataset.

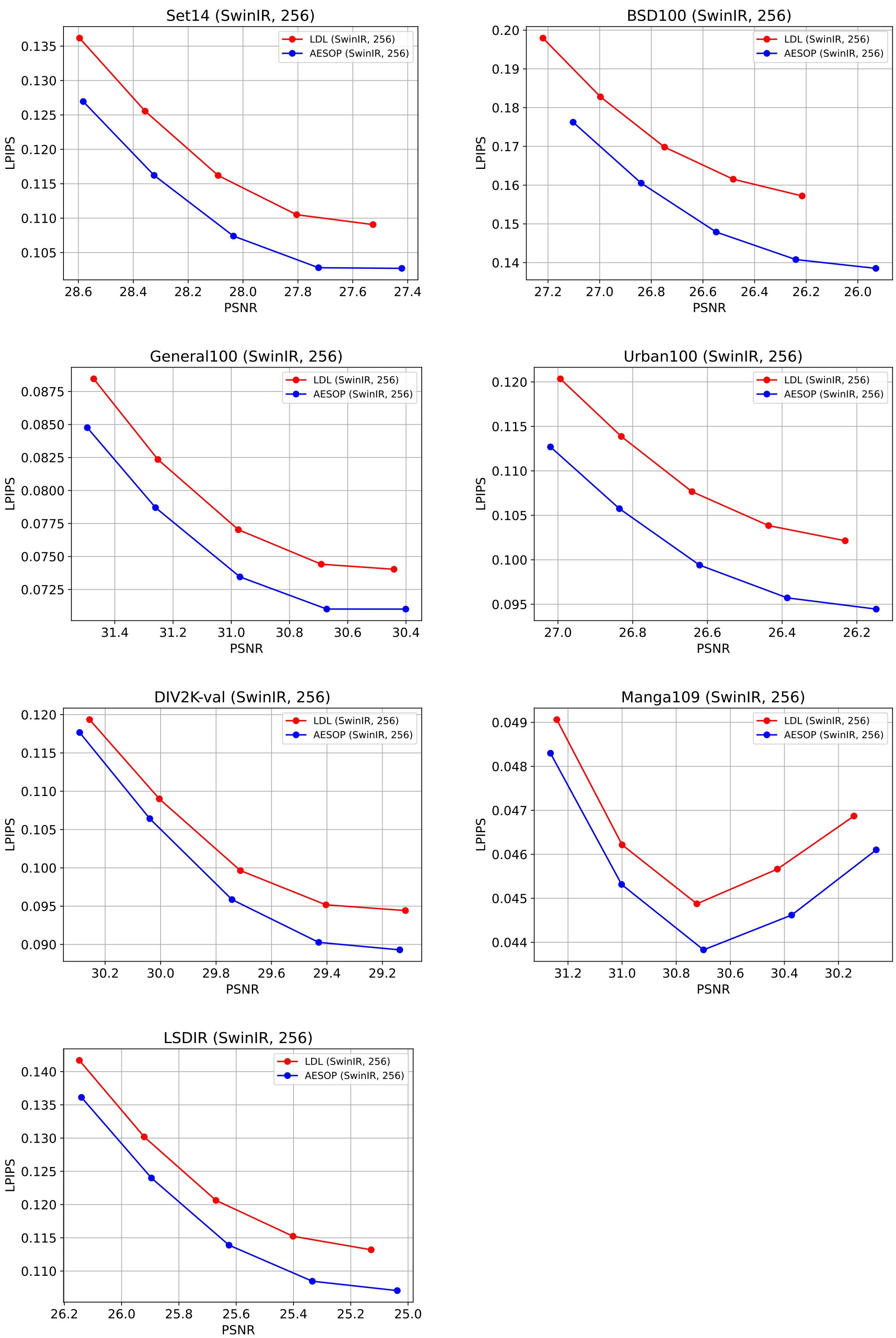


Figure 18. The perception-distortion trade-off curve between AESOP and baseline methods on top of the SwinIR [36] backbone. The training HR patch size is 256. AESOP often fails to improve the performance on the Manga109 dataset.

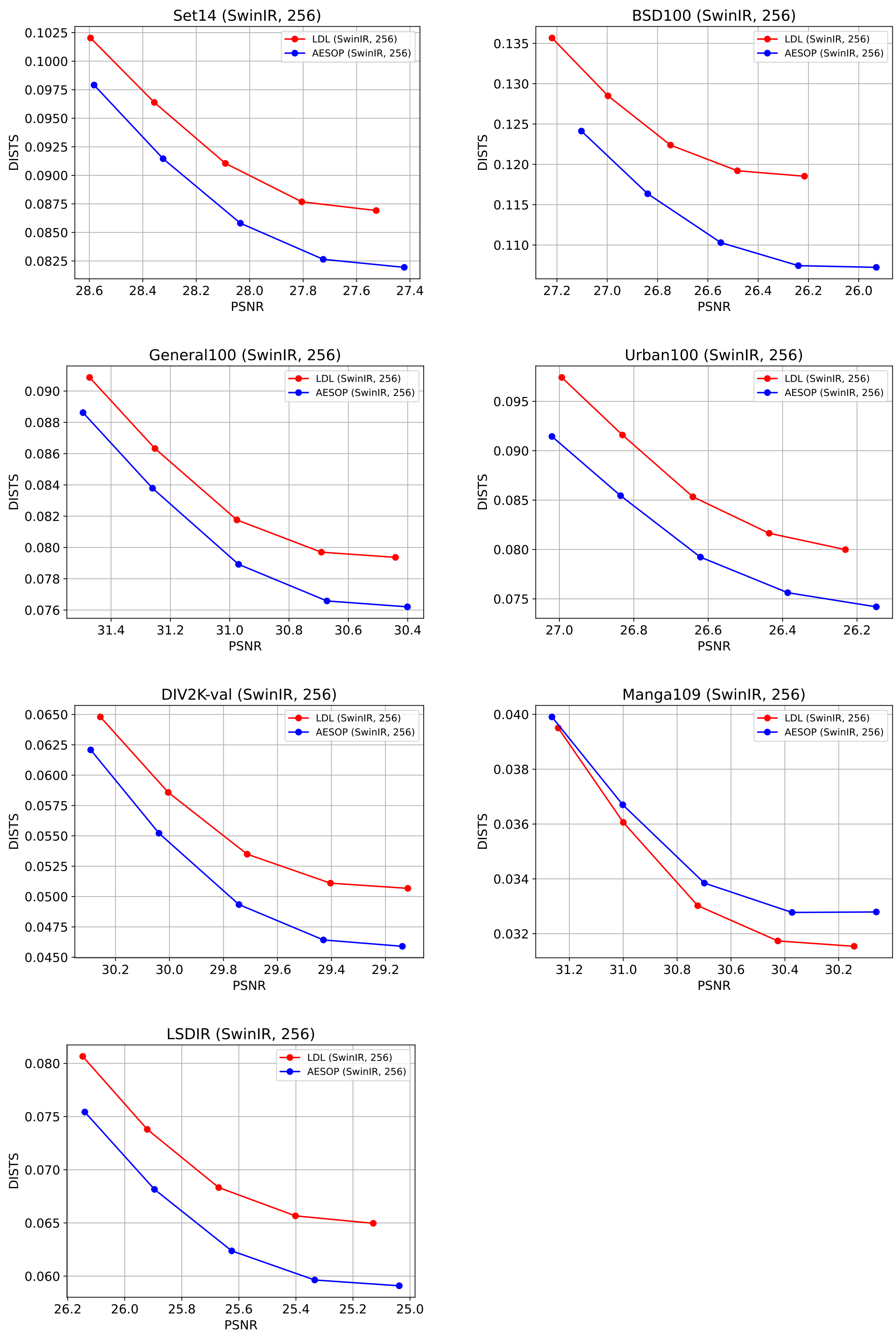


Figure 19. The perception-distortion trade-off curve between AESOP and baseline methods on top of the SwinIR [36] backbone. The training HR patch size is 256. AESOP often fails to improve the performance on the Manga109 dataset.

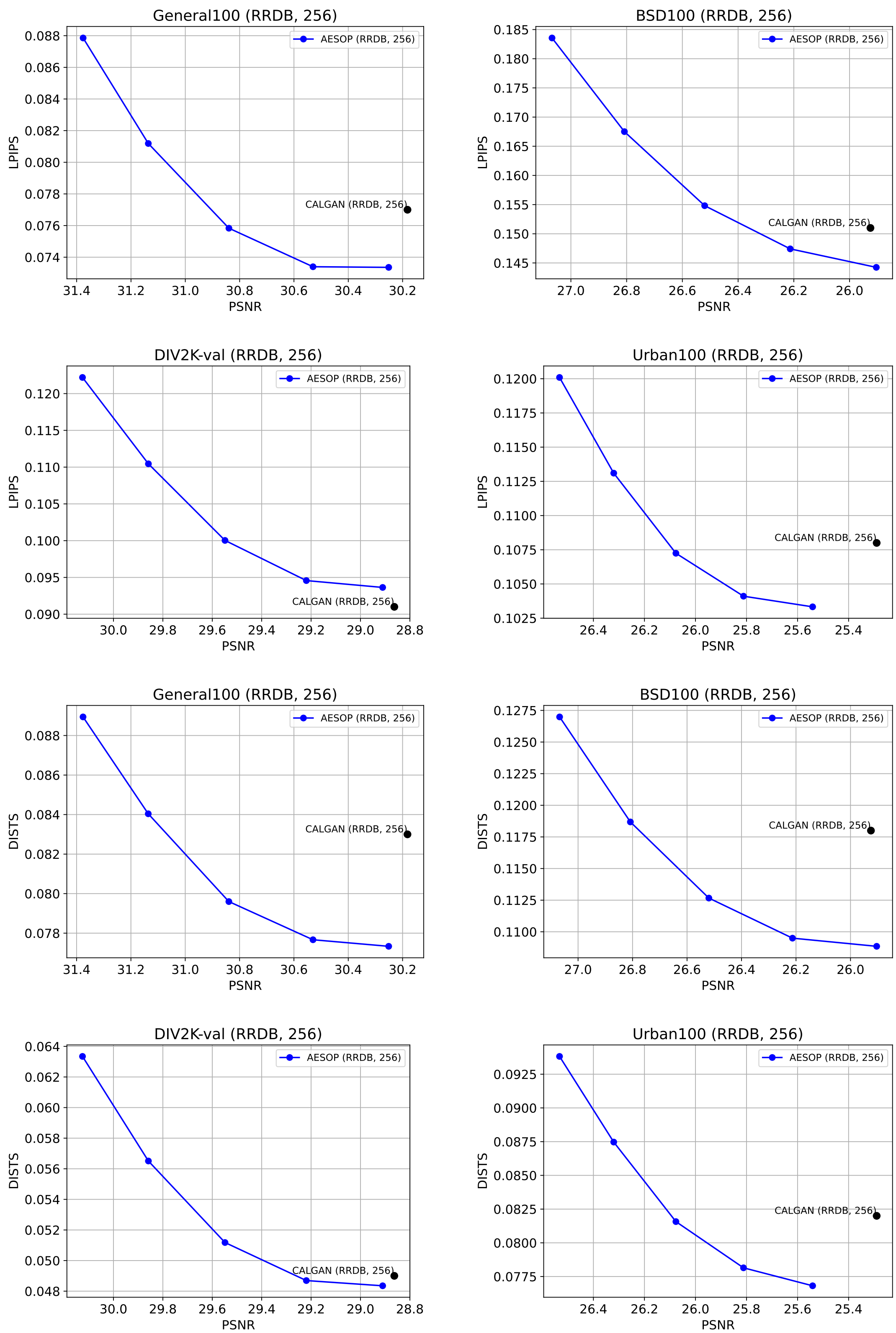


Figure 20. The perception-distortion trade-off curve between AESOP and baseline methods on top of the RRDB [54] backbone. The training HR patch size is 256. AESOP mostly outperforms CALGAN [48] even without the MoE-discriminator.

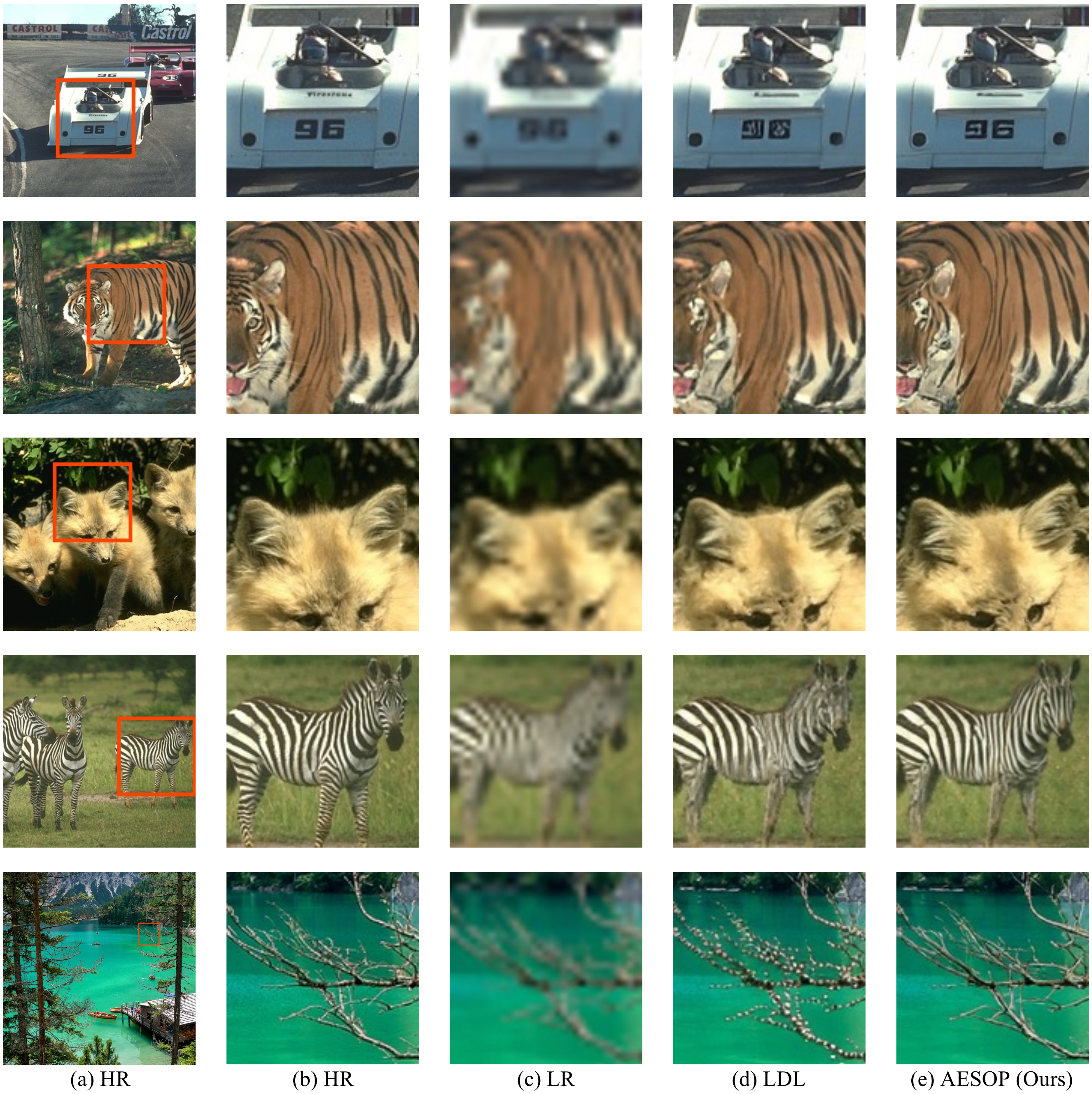
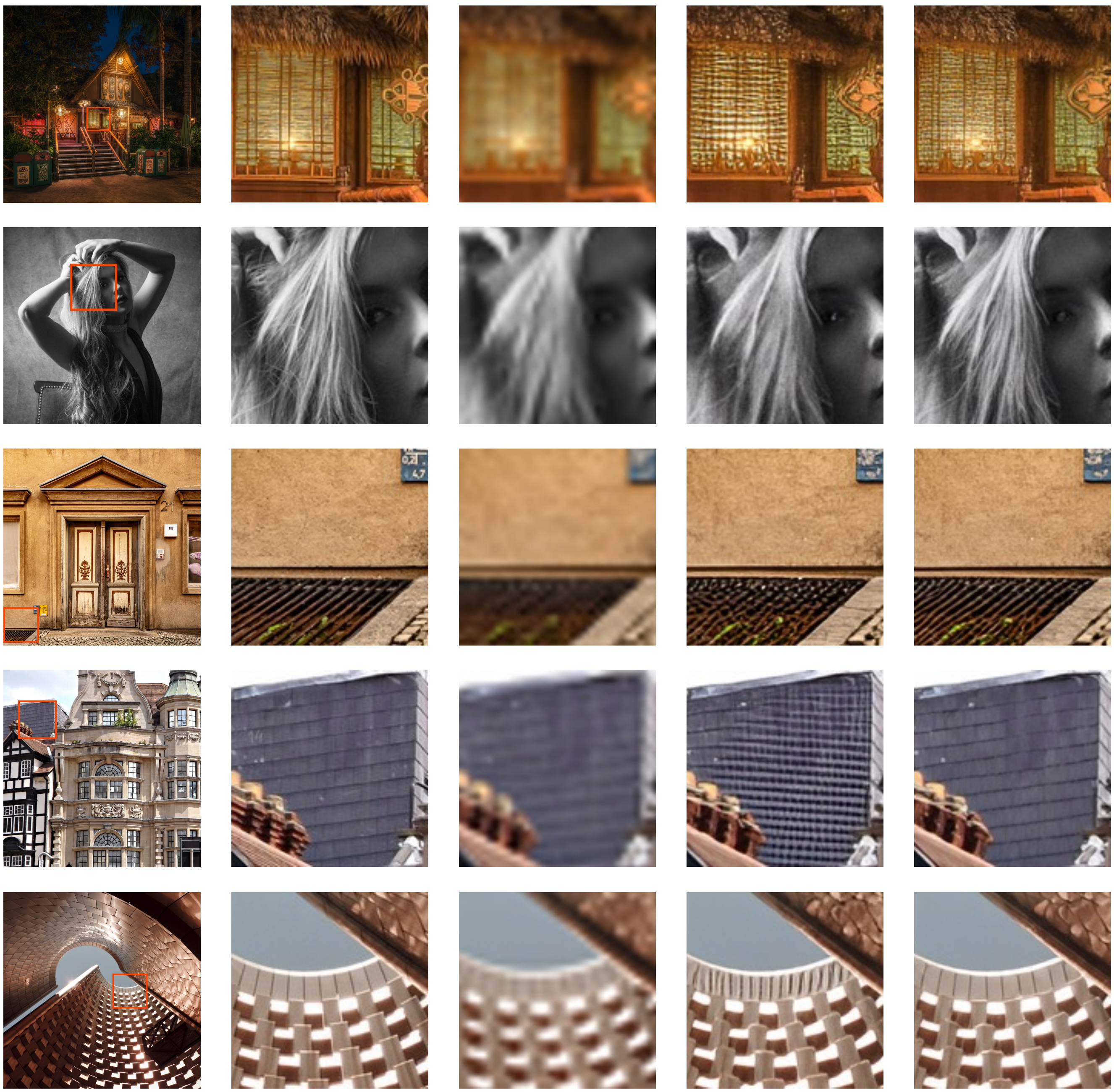
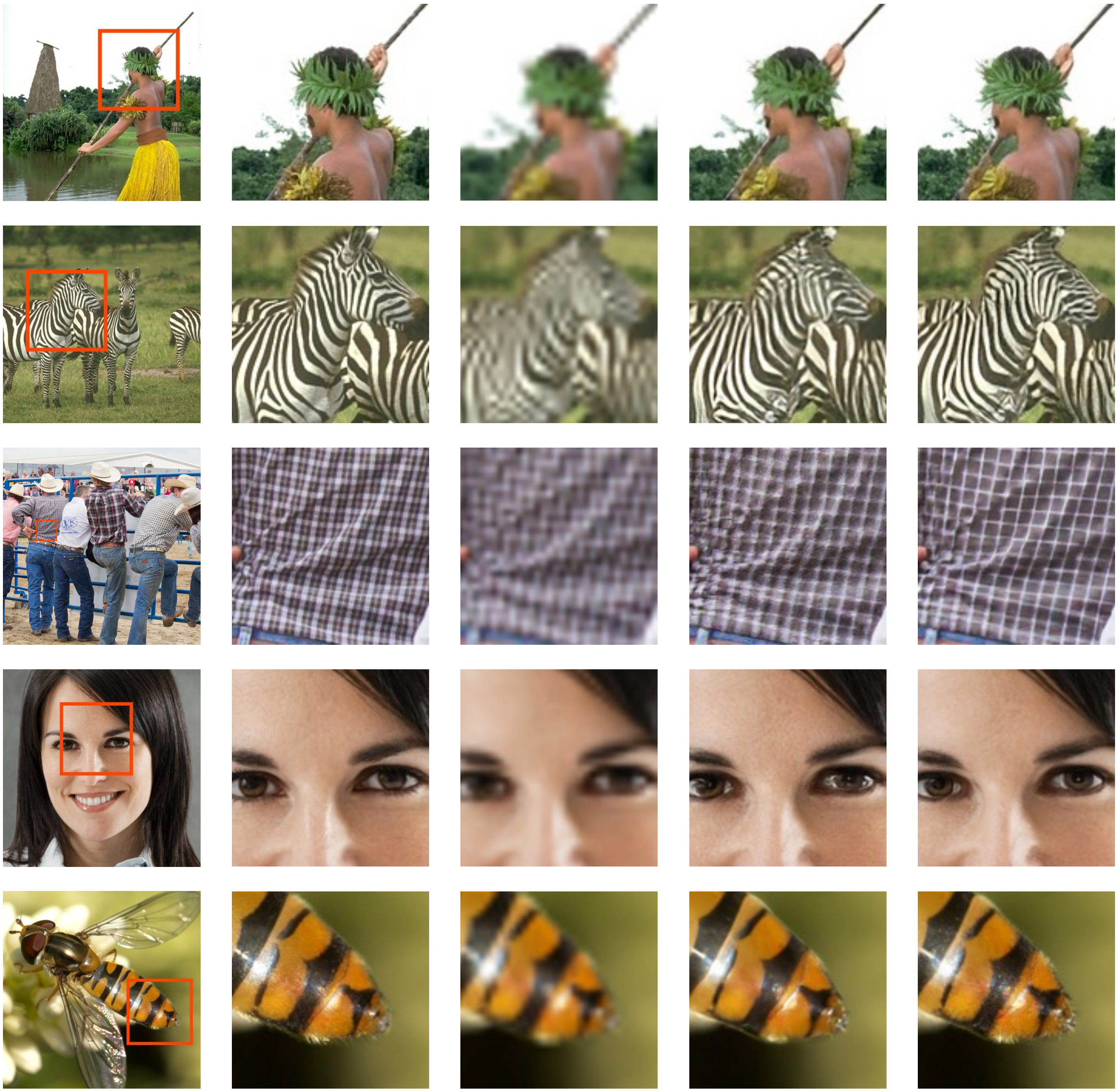


Figure 21. Visualization of AESOP (ours) and the baseline method for the bicubic ×4 SR task with SwinIR backbone. AESOP can generate fine-grained textures with a lower level of visual artifacts. **Zoom in for best view**.



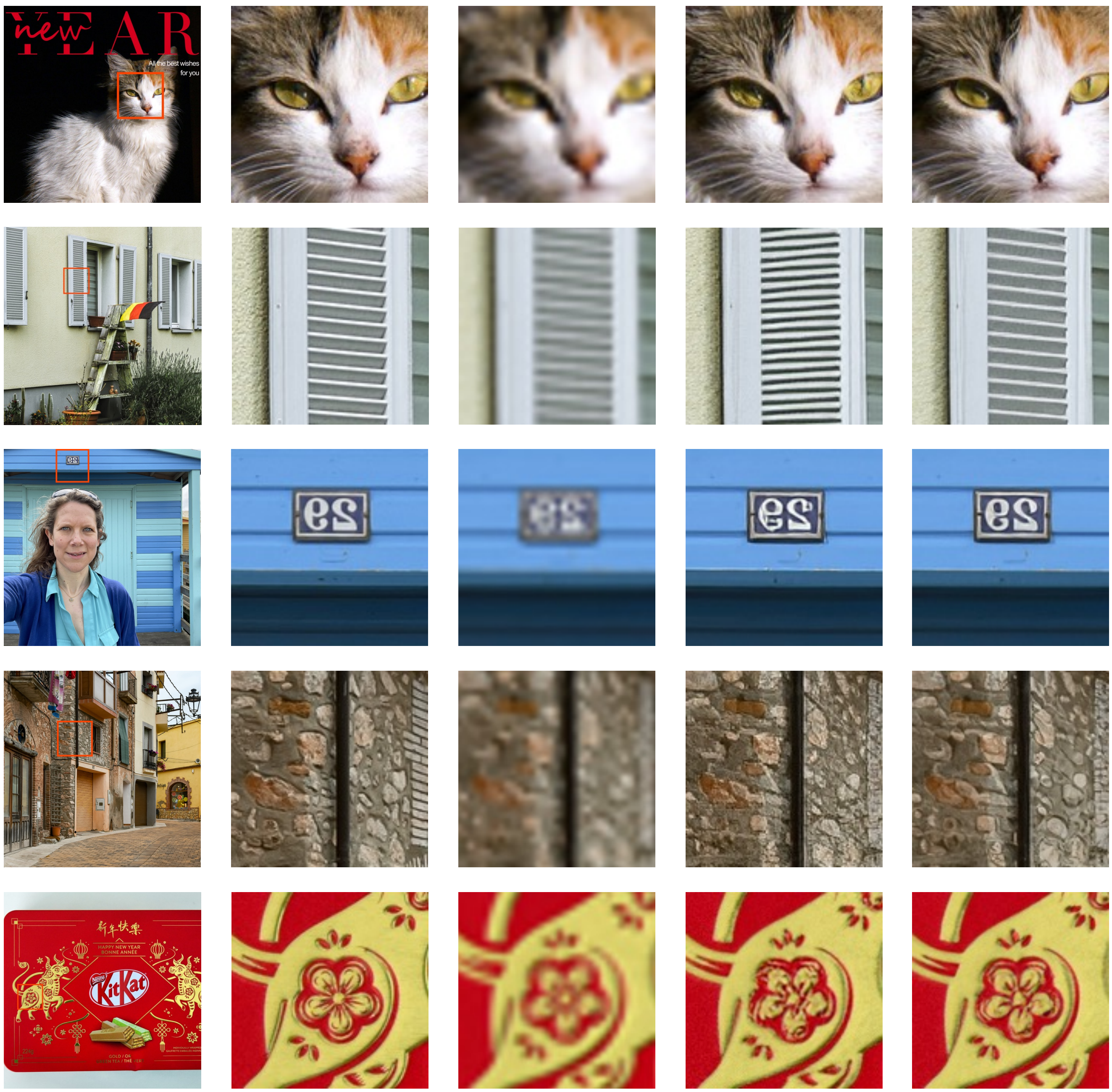
(a) HR (b) HR (c) LR (d) LDL (e) AESOP (Ours)

Figure 22. Visualization of AESOP (ours) and the baseline method for the bicubic ×4 SR task with SwinIR backbone. AESOP can generate fine-grained textures with a lower level of visual artifacts. **Zoom in for best view**.



(a) HR (b) HR (c) LR (d) LDL (e) AESOP (Ours)

Figure 23. Visualization of AESOP (ours) and the baseline method for the bicubic ×4 SR task with RRDB backbone. AESOP can generate fine-grained textures with a lower level of visual artifacts. **Zoom in for best view**.



(a) HR (b) HR (c) LR (d) LDL (e) AESOP (Ours)

Figure 24. Visualization of AESOP (ours) and the baseline method for the bicubic ×4 SR task with RRDB backbone. AESOP can generate fine-grained textures with a lower level of visual artifacts. **Zoom in for best view**.