Auto-Encoded Supervision for Perceptual Image Super-Resolution

MinKyu Lee, Sangeek Hyun, Woojin Jun, Jae-Pil Heo*
Sungkyunkwan University
{bluelati98, hsi1032, junwoojin, jaepilheo}@skku.edu

Abstract

*This work tackles the **fidelity** objective in the **perceptual** super-resolution (SR) task. Specifically, we address the shortcomings of pixel-level $\mathcal{L}_p$ loss ($\mathcal{L}_{pix}$) in the GAN-based SR framework. Since $\mathcal{L}_{pix}$ is known to have a trade-off relationship against perceptual quality, prior methods often multiply a small scale factor or utilize low-pass filters. However, this work shows that these circumventions fail to address the fundamental factor that induces blurring. Accordingly, we focus on two points: 1) precisely discriminating the subcomponent of $\mathcal{L}_{pix}$ that contributes to blurring, and 2) only guiding based on the factor that is free from this trade-off relationship. We show that they can be achieved in a surprisingly simple manner, with an Auto-Encoder (AE) pretrained with $\mathcal{L}_{pix}$. Accordingly, we propose the Auto-Encoded Supervision for Optimal Penalization loss ($\mathcal{L}_{AESOP}$), a novel loss function that measures distance in the **AE space** [1], instead of the raw pixel space. By simply substituting $\mathcal{L}_{pix}$ with $\mathcal{L}_{AESOP}$, we can provide effective reconstruction guidance without compromising perceptual quality. Designed for simplicity, our method enables easy integration into existing SR frameworks. Extensive experiments demonstrate the effectiveness of AESOP.*

1. Introduction

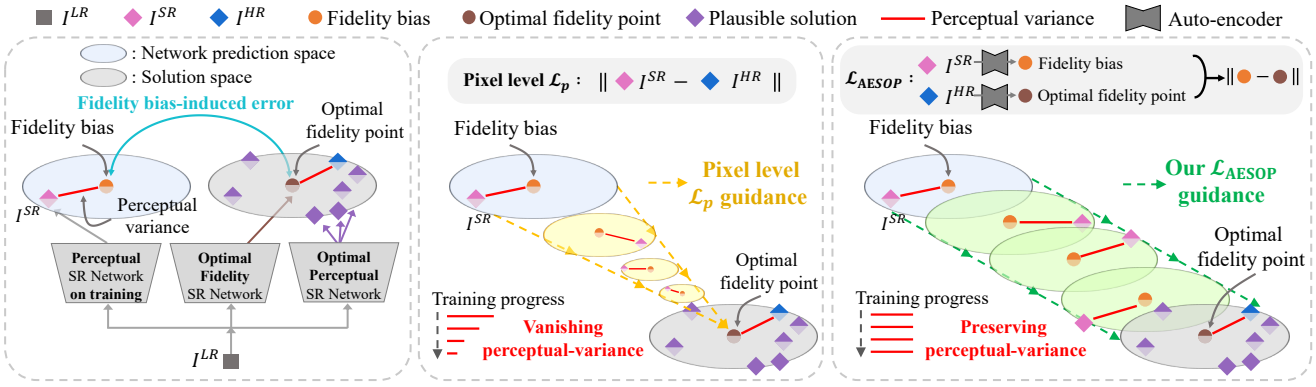
Image Super-Resolution is a fundamental challenge in image processing, where the goal is to reconstruct an unknown high-resolution (HR) image from its low-resolution (LR) counterpart. Recent advances in this field have branched into two distinct mainstreams; fidelity-oriented SR and perceptual quality oriented SR (perceptual SR). Fidelity-oriented SR methods aim for high fidelity towards the HR image on a pixel-wise basis. These methods generally adopt the per-pixel reconstruction loss $\mathcal{L}_{\text{pix}}$ (i.e., pixel-level $\mathcal{L}_{\text{p}}$ loss), thereby regressing the unique point that minimizes the expected error. This unique point of minimum expected error is known as the average point over multiple plausible solutions [31], which we will refer to as the *optimal fidelity point* (brown dot in Fig.1.(a)) throughout this work. Another branch of research is the perceptual SR task where the emphasis is on generating visually plausible SR images rather than mere minimum pixel-wise error. Notably, the inherent ill-posedness leads perceptual SR to exhibit a range of variant realistic solutions (multiple purple points in Fig.1.(a)), each pivotal to the aforementioned optimal fidelity point.

Here, the representative framework in perceptual SR is the SRGAN-based [30] framework, which utilizes perceptual quality oriented losses [22, 54] together with $\mathcal{L}_{\text{pix}}$, the *de facto* training scheme. Yet, since these methods rely on $\mathcal{L}_{\text{pix}}$ as their fidelity loss term, they cannot avoid the blurring phenomenon as shown in the perception-distortion (PD) trade-off [2]. To address this, they either apply a small coefficient [30, 54] to $\mathcal{L}_{\text{pix}}$ or use low-pass filters (LPF) [11, 69] before calculating $\mathcal{L}_{\text{pix}}$. However, we point out that these circumventions result in suboptimal performance, as they misinterpret the implications of $\mathcal{L}_{\text{pix}}$ and fail to distinguish between factors that cause blurring and those that do not.

Accordingly, this work revisits $\mathcal{L}_{\text{pix}}$ and aims to correctly analyze the implications of it, by introducing two key factors of an SR image: 1) the *perceptual variance* factor, and 2) the *fidelity bias* factor. Here, the perceptual variance is a necessary variance that captures realistic textures and fine details (red line in Fig.1.(a)). Meanwhile, the fidelity bias is the residual component of the SR image, apart from the perceptual variance factor. This can be understood as the blurry average solution without the fine-grained texture that possesses randomness (no variance, Fig.3.(d)), or can also be understood as the centroid of the distribution where the SR image was originally expected to be sampled from (orange dot in Fig.1.(a)). For an optimal SR image (or the HR), the fidelity bias of itself is the optimal fidelity point.

In terms of these two key components defined above, we will show that $\mathcal{L}_{\text{pix}}$ is identical to minimizing both the fidelity bias induced errors and the perceptual variance factor. However, while minimizing the fidelity bias induced error is an intended aspect of $\mathcal{L}_{\text{pix}}$, vanishing perceptual variance is not suitable in perceptual SR. Specifically, when the perceptual variance factor is minimized, the prediction space degenerates and the SR image converges to the blurry average

*Corresponding author.
[1] AE space indicates the space *after* the decoder, not the bottleneck.



(a) General SR system **(b) Pixel-level $\mathcal{L}_p$ guidance** **(c) $\mathcal{L}_{AESOP}$ (Ours) reconstruction guidance**

Figure 1. Conceptual illustration of the proposed AESOP loss and the pixel-level $\mathcal{L}_p$ reconstruction guidance employed in typical perceptual SR methods. **(a)** Fidelity oriented SR network trained with $\mathcal{L}_{pix}$ estimates the average over plausible solutions (i.e., the optimal fidelity point). Meanwhile, perceptual SR involves a range of multiple solutions, standing around the optimal fidelity point. Thus, we identify two fundamental components of a perceptual SR image as 1) the perceptual variance factor (red line), a factor that possesses randomness and contributes to realistic textures, and 2) the fidelity bias term (orange dot), the residual blurry component of an SR image, contributing to the overall fidelity, apart from the perceptual variance. **(b)** Typical perceptual SR methods adopt $\mathcal{L}_{pix}$ for reconstruction guidance, which pushes the perceptual variance factor to vanish. Thus, when combined with perceptual quality oriented losses that encourage this variance factor, conflict arises, leading to suboptimal performance. **(c)** In contrast, $\mathcal{L}_{AESOP}$ only penalizes the fidelity bias-induced error, while preserving these critical perceptual variance factors. This ensures improved fidelity without sacrificing perceptual quality.

image as in Fig.1.(b). Meanwhile, perceptual quality oriented losses aim to preserve this necessary perceptual variance, which indicates a conflict against $\mathcal{L}_{pix}$. Consequently, as long as the SR network receives reconstruction guidance from $\mathcal{L}_{pix}$, the perceptual quality oriented losses cannot converge to an optimal point, limiting the visual quality.

Now, we focus on the other counterpart of the SR image, the fidelity bias, and the error induced by it. The fidelity bias induced error indicates the overall degree of misalignment between the prediction space and the solution space. Accordingly, reducing the fidelity bias induced error is identical to aligning the centroids of the prediction space and the solution space, without altering the range of the prediction space (i.e., preserving perceptual variance), as in Fig.1.(c). Thus, by reducing the fidelity bias induced error, we can achieve improved fidelity without degrading perceptual variance, which is a highly desired aspect for optimal reconstruction guidance in the perspective of perceptual SR.

Accordingly, this motivates us to design a novel reconstruction loss that can precisely discriminate the fidelity bias and the perceptual variance, and solely penalize based on the fidelity bias term. Notably, we will show that this can be surprisingly simplified with a pretrained Auto-Encoder (AE). We will pretrain an AE with $\mathcal{L}_{pix}$, and take $\mathcal{L}_p$ in the AE space instead of raw pixel space. Since $\mathcal{L}_{pix}$ removes perceptual variances, loss in the AE space will enable us to penalize solely based on fidelity biases. Importantly, we are conversely taking advantage of the vanishing perceptual variance phenomenon of $\mathcal{L}_{pix}$, which was observed as a critical limitation. We refer to this as the Auto-Encoded Supervision for Perceptual SR (AESOP).

In summary, our contributions can be simplified as follows. First, we point out the implications of $\mathcal{L}_{pix}$ in the context of the perceptual SR task, which has often been misunderstood. Second, we propose a novel reconstruction loss that only penalizes the fidelity bias factors, thereby preserving visually important perceptual variance factors of SR images. Finally, we provide extensive experiments that validate the effectiveness of AESOP, leading to significant improvement in the perceptual SR task.

Disclaimer. Perceptual oriented losses are responsible for handling the preserved perceptual variance term, thereby generating realistic textures and fine-details. We keep improvements on these losses out of the scope of this work.

2. Related work

Fidelity-oriented SR. The pioneering works [13, 25], CNN-based [9, 16, 39, 45, 47, 68, 70] and Swin Transformer [40] based methods [5, 6, 34, 36, 63] have shown remarkable improvements. The majority of these methods employ the per-pixel loss $\mathcal{L}_{pix}$ as their sole objective. This leads them to estimate the average over multiple solutions [20, 41], resulting in high PSNR scores but is empirically shown to be blurry. In our work, the *optimal fidelity point* is the optimal estimation of these fidelity-oriented SR methods.

Perceptual SR. The emphasis here is on visual quality over mere per-pixel error. These methods [30, 42, 64, 67] commonly adopt the SRGAN [30]-based framework, integrating $\mathcal{L}_{pix}$ with perceptual-quality-oriented losses such as perceptual loss [22] and adversarial loss [15]. This framework aims to enhance visual quality while retaining a fair amount

of fidelity to the HR image. Recently, diffusion-based SR methods [50, 53, 61] have shown significant progress. However, due to their limitations in SR tasks without complex degradations and the high computational cost, GAN-based SR remains as one of the main branches of research [29]. We limit the scope of our work to GAN-based SR methods.

Improvements in the SRGAN-framework. Advancements were made on the adversarial loss [27, 54], discriminator architecture [33, 48] and the perceptual loss [10, 28, 49], and also in enhancing GAN stability [27, 37, 42, 59]. Despite remarkable improvements, all these efforts primarily concentrate on perceptual-oriented loss factors. Meanwhile, the use of $\mathcal{L}_{\text{pix}}$ for perceptual SR tasks has not been thoroughly investigated. To the best of our knowledge, this is the first attempt that successfully tackles the *fidelity*-loss in a *perceptual* SR framework, removing the problematic pixel-level reconstruction loss in the SRGAN-framework.

3. Revisiting per-pixel loss in perceptual SR

Considering the Oracle case. Consider an optimal perceptual SR network that can sample images from the true posterior. By construction, SR images generated from this network are valid solutions but are not necessarily a pixel-wise exact match to the specific HR image instance in our dataset, due to the inherent ill-posed nature of SR. Yet, $\mathcal{L}_{\text{pix}}$ compares two images on a strict pixel-wise basis. This results in penalizing the SR image, despite it being an ideal solution in the context of perceptual SR, by construction.

Revisiting $\mathcal{L}_{\text{pix}}$ in perceptual SR. The phenomenon where even an optimal network gets penalized under $\mathcal{L}_{\text{pix}}$ results in blurred texture. This is grounded in the fact that training with $\mathcal{L}_{\text{pix}}$ is effectively a Maximum Likelihood Estimation, which jointly minimizes both bias errors and also *variance in predictions*. Specifically, prior works [21, 32] have shown that this can be decomposed into jointly minimizing two terms: the systematic-effect (SE) term and the variance-effect (VE) term, which are induced by the bias and variance of the prediction, respectively. Formally, given a symmetric loss function $\mathcal{L}$ and $y \sim p(y|x)$ for an input x, the training objective $\min_{\hat{y}} \mathbb{E}_{y,\hat{y}}[\mathcal{L}(y,\hat{y})]$ can be simplified [2] as:

$$\min_{\hat{y}}\{\underbrace{\mathbb{E}_y[\mathcal{L}(y,\mu_{\hat{y}}) - \mathcal{L}(y,\mu_y)]}_{\text{SE}(y,\hat{y})\text{: LF + regressable HF}} + \underbrace{\mathbb{E}_{y,\hat{y}}[\mathcal{L}(y,\hat{y}) - \mathcal{L}(y,\mu_{\hat{y}})]}_{\text{VE}(y,\hat{y})\text{: non-regressable HF}}\}, \tag{1}$$

with $\hat{y}$ as an estimator of y, and $\mu_y = \arg\min_\mu \mathbb{E}_y[\mathcal{L}(y,\mu)]$ and $\mu_{\hat{y}} = \arg\min_\mu \mathbb{E}_{\hat{y}}[\mathcal{L}(\hat{y},\mu)]$. For $\mathcal{L}_2$, the two terms above are further simplified as follows:

$$\begin{aligned}\text{SE}(y,\hat{y}) &= \mathbb{E}_y[(y-\mu_{\hat{y}})^2 - (y-\mu_y)^2] = (\mu_{\hat{y}} - \mu_y)^2 \\ \text{VE}(y,\hat{y}) &= \mathbb{E}_{y,\hat{y}}[(y-\hat{y})^2 - (y-\mu_{\hat{y}})^2] = \mathbb{E}_{\hat{y}}[(\hat{y}-\mu_{\hat{y}})^2].\end{aligned} \tag{2}$$

[2] The irreducible variance term of y is omitted here.

SE and VE in terms of perceptual SR. For the perceptual SR task, SE minimization is desired but VE should be sufficiently preserved. We elaborate on the details below.

VE refers to the additional error introduced by generating components with inherent randomness. In perceptual SR, this is a necessary and inevitable error term induced by fine-grained textures, which cannot be learned via regression. Accordingly, we define the VE term as the *perceptual variance* factor. Here, minimizing VE can be understood as further reducing the expected pixel-wise error, apart from reducing SE, at the cost of removing visually important fine textures. In other words, VE is the factor that leads to the PD trade-off due to its randomness, and VE minimization pushes the prediction space to degenerate, only accepting the average solution as in Fig.1.(b), which is blurry [30].

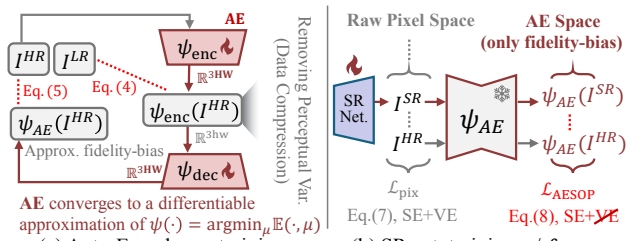
Focusing now on SE, this is an *unnecessary error term* that can be reduced without inducing PD trade-off since it does not have randomness. Intuitively, it is the overall degree of alignment between the SR and HR image. More specifically, this is the distance between the centroids of the two distributions, where the HR and SR images are each expected to be sampled from. Since these centroids are the minimum expected error points of each distribution, they resemble the fidelity-oriented SR counterpart of the perceptual SR images [31], as in Fig.1.(a). Accordingly, we define each factor in the SE term (μ_y, $\mu_{\hat{y}}$) as the *fidelity biases*, and SE itself as the fidelity bias induced error. Note that fidelity biases are not simply low-frequency (LF) components. As often reflected in fidelity-oriented SR methods, specific HF components such as simple object boundaries and edges can be learned via pixel-level regression. Fidelity biases include these regressable high-frequency components.

Overall, since VE is the sole factor inducing the PD trade-off, it is straightforward that we can safely reduce SE without harming perception. By minimizing SE for a given VE, we can obtain the maximum fidelity for a given perception level; the ideal PD trade-off [3]. At the same time, a sufficient level of VE should be preserved for visual quality. Accordingly, the following sections will elaborate on designing a novel loss that minimizes SE while preserving VE, taking a step toward the optimal perceptual SR network.

Revisiting prior methods. Observations above share some key concepts with the well-known perception-distortion (PD) trade-off [2]. However, we highlight aspects of $\mathcal{L}_{\text{pix}}$ that are often overlooked in most training methods, with the terms defined above. Specifically, previous approaches aim to avoid blurring by either 1) introducing LPF before loss calculation [11, 69] or 2) simply applying a small coefficient for $\mathcal{L}_{\text{pix}}$ [30, 37, 54]. Below, we will show that both are misinterpreting the blurring phenomenon.

First, LPF-based approaches remove more information

[3] Zero SE alone does not indicate optimal *perception*.



(a) Auto-Encoder pretraining (b) SR net. training w/ $\mathcal{L}_{\text{AESOP}}$

Figure 2. **(a)** We pretrain an Auto-Encoder ψ_{AE} that removes perceptual variance factors, thereby establishing a feature space where only the fidelity bias factors reside. **(b)** The main SR network training step with the proposed $\mathcal{L}_{\text{AESOP}}$. By applying reconstruction objectives such as the $\mathcal{L}_1$ loss in the auto-encoded space, we can solely target the fidelity bias induced error without suffering from vanishing perceptual variance (i.e., suffer from blurring). We omit perceptual-quality-oriented losses here.

than required. They assume that the fidelity biases solely consist low-frequency image components. Yet, bias factors $(\mu_y, \mu_{\hat{y}})$ include certain high-frequency components as discussed above. Therefore, despite that LPF-based approaches can avoid texture blurring induced by VE, they also fail to guide regressable high-frequency components that are free from texture blurring. By not providing guidance for certain components of SE, they achieve lower fidelity than necessary, failing to reach the optimal PD trade-off. We provide further analysis for this in Sec.5.2.

Meanwhile, applying a small coefficient to $\mathcal{L}_{\text{pix}}$ misguidedly treats all aspects of $\mathcal{L}_{\text{pix}}$ as contributing to blurring. This indiscriminative approach unintentionally weakens SE reduction. When combined with the adversarial loss, this also leads to redundant SE, failing to reach the optimal PD trade-off. This is because adversarial loss works in a task-blind, unsupervised manner: it improves realism but does not consider the alignment between the input image and the network output. As this significantly hinders SE convergence [37, 59], strong guidance on the SE factor is required to prevent high SE. However, the small coefficient greatly reduces this guidance, resulting in unnecessary fidelity loss and a suboptimal PD trade-off. Additionally, since $\mathcal{L}_{\text{pix}}$ fundamentally enforces VE reduction, it can also never achieve optimal perception despite applying a small scale factor. See Appendix.E for discussions and graphical illustrations.

4. Method

Motivation. An optimal perceptual SR should minimize fidelity bias induced errors while preserving perceptual variance. However, prior methods often fail to achieve this due to their inability to effectively discriminate these factors. This motivates us to design a new method that disentangles them, and solely penalizes based on fidelity biases, bringing us closer to the optimal perceptual SR.

Overview. The proposed method can be simplified into two steps. First, we develop an Auto-Encoder (AE), tailored to create a feature space exclusively for fidelity biases. Second, instead of taking $\mathcal{L}_{\text{p}}$ in raw pixel space as typical methods, we calculate $\mathcal{L}_{\text{p}}$ in the AE space as in Fig.2. Taking $\mathcal{L}_{\text{p}}$ in the AE space enables us to provide effective reconstruction guidance with preserved perceptual variance. Note that the term *AE space* indicates the space after the decoder, not the bottleneck. Fundamentally, we are utilizing an AE as a differentiable approximation of an operator $\psi(\cdot) := \arg\min_\mu \mathbb{E}[\mathcal{L}(\cdot, \mu)]$ to substitute $\mathcal{L}_{\text{pix}}$ with SE.

Baseline. This work aims to make improvements in the GAN-based perceptual SR task. Accordingly, we follow a recent GAN-based SR method LDL [37], and set our baseline training objective as below:

$$\mathcal{L}_{\text{base}} = \lambda_1 \mathcal{L}_{\text{pix}} + \lambda_2 \mathcal{L}_{\text{percep}} + \lambda_3 \mathcal{L}_{\text{adv}} + \lambda_4 \mathcal{L}_{\text{artif}}, \quad (3)$$

where $\mathcal{L}_{\text{pix}}$, $\mathcal{L}_{\text{percep}}$, $\mathcal{L}_{\text{adv}}$, $\mathcal{L}_{\text{artif}}$ are the widely used pixel-level $\mathcal{L}_{\text{p}}$ loss, perceptual loss [22], the adversarial loss [54], and the artifact loss [37], respectively, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are coefficients for each loss factors, respectively. We limit the scope of this work to tackling the *fidelity* loss term of *perceptual* SR, thus, we will only modify $\mathcal{L}_{\text{pix}}$, while leaving all other loss terms unchanged.

4.1. Auto-Encoder pretraining

Designing the fidelity bias feature space. Our aim is to provide reconstruction supervision focused exclusively on SE. Given that the only components of SE are the fidelity biases $(\mu_y, \mu_{\hat{y}})$, our task simplifies into estimating an operator ψ that estimates the fidelity bias of a given image as $\psi(\cdot) := \arg\min_\mu \mathbb{E}[\mathcal{L}(\cdot, \mu)]$. To maintain simplicity, we employ a basic Auto-Encoder (AE) to construct a differentiable approximation of this operator. This architectural choice is grounded by the nature of SR, where the HR images are conditioned by the LR images and the scale factor. Thus, we model this relationship as $y \sim p(y|x)$ with $x \equiv \phi(y, s)$, where ϕ is the $\times s$ downsampling function. Additionally, the definition of ψ involves minimizing the expected loss over this conditional distribution. Thus, we pretrain the AE with $\mathcal{L}_p$ to learn the forward mapping $x \leftarrow y$, and consecutively, the inverse mapping $y \leftarrow x$. Now, the pretrained AE will act as a differentiable approximation of ψ, which can decompose the fidelity bias of images and can also be directly plugged into the training framework.

To provide further intuition, we emphasize that the bottleneck of our AE is designed to have the same dimensionality as the LR image. Contrary to most AEs or feature encoders [22], which use a high-dimensional latent space to learn *additional semantics or high-level representations* beyond the raw pixel space, our AE is specifically designed to *remove* particular *low-level* features from the pixel space. The carefully chosen architecture and pretraining objective form an information bottleneck that effectively compresses out factors that have inherent randomness. Since this is the

(a) HR (b) LDL (c) AESOP (Ours) (d) AutoEncoded (e) $\mathcal{L}_{\text{pix}}$ (f) $\mathcal{L}_{\text{AESOP}}$

Figure 3. Key components of $\mathcal{L}_{\text{AESOP}}$ and $\mathcal{L}_{\text{pix}}$ on SwinIR-backbone. $\mathcal{L}_{\text{pix}}$ in (e) penalizes perceptually-variance factors, leading to blurry images in (b). In contrast, $\mathcal{L}_{\text{AESOP}}$ in (f) only penalizes based the fidelity bias (d), which enables us to obtain increased realism as in (c).

perceptual variance, we can isolate only the fidelity bias as in Fig.2. Thus, $\mathcal{L}_p$ in the AE space resembles loss between fidelity biases, which is fundamentally identical to SE: a term that improves fidelity without inducing blurring. Notably, while vanishing perceptual variance was observed as a critical limitation of $\mathcal{L}_{\text{pix}}$ in the perceptual SR task, we are conversely taking advantage of it by removing perceptual variance factors in our newly designed fidelity loss term.

AE pretraining. To design a fidelity bias estimator, we pretrain our AE to approximate $\psi(\cdot) := \arg\min_{\mu} \mathbb{E}[\mathcal{L}(\cdot, \mu)]$ for $y \sim p(y|x)$ with $x \equiv \phi(y, s)$. Thus, the AE consecutively estimates the low-resolution counterpart x and then reconstructs y. Accordingly, the pretraining objective is straightforward as follows:

$$\mathcal{L}^{\text{rec}}_{\text{LR}} = ||\psi_{\text{enc}}(I^{\text{HR}}) - I^{\text{LR}}||_p \quad (4)$$

$$\mathcal{L}^{\text{rec}}_{\text{HR}} = ||\psi_{\text{AE}}(I^{\text{HR}}) - I^{\text{HR}}||_p, \quad (5)$$

where $\psi_{\text{AE}} := \psi_{\text{dec}} \cdot \psi_{\text{enc}}$ denotes the AE, ψ_{enc}, ψ_{dec} is the encoder and decoder, and I^{LR}, I^{HR} are LR, HR images. The AE will act as an effective bias estimator, enabling us to design a space where only fidelity biases reside. Note that these losses are only used to pretrain the AE, and will not be used when training the SR network.

AE architecture. Based on the constructions above, the encoder takes an HR image, and estimates the corresponding LR versions; and the decoder, vice-versa as follows:

$$\psi_{\text{enc}} := \mathbb{R}^{3\text{HW}} \mapsto \mathbb{R}^{3\text{hw}}, \quad \psi_{\text{dec}} := \mathbb{R}^{3\text{hw}} \mapsto \mathbb{R}^{3\text{HW}}, \quad (6)$$

where HW and hw each indicate the spatial dimension of the HR and LR images. Since the decoding process resembles a fidelity oriented SR task, we employ an off-the-shelf SR architecture RRDBNet [54], and initialize the decoder as the pretrained weights for the fidelity-oriented SR task. The encoder is simply a lightweight CNN with downsampling. Refer to the appendix for further details.

Bottleneck collapse. Consider a scenario where the encoder exactly matches the corresponding LR image of the input. If the SR image simply downscales to the original LR image, no loss would backpropagate regardless of the regressable high-frequency component quality of the SR image. Since this can potentially harm the performance, the encoder is jointly optimized with the decoder for Eq.(5).

4.2. Auto-Encoded supervision

Defining the AESOP loss. Since we have obtained a feature space that only retains the fidelity bias factors, we finally define $\mathcal{L}_{\text{AESOP}}$ as $\mathcal{L}_{\text{p}}$ with auto-encoded versions of HR and SR images. Contrary to $\mathcal{L}_{\text{pix}}$ which minimizes both SE and VE, the proposed $\mathcal{L}_{\text{AESOP}}$ only minimizes SE is as follows:

$$\mathcal{L}_{\text{pix}} = ||I^{\text{HR}} - I^{\text{SR}}||_p \quad (= \text{SE} + \text{VE}), \quad (7)$$

$$\mathcal{L}_{\text{AESOP}} = ||\psi_{\text{AE}}(I^{\text{HR}}) - \psi_{\text{AE}}(I^{\text{SR}})||_p \quad (\approx \text{SE} + \cancel{\text{VE}}), \quad (8)$$

where I^{HR}, I^{SR} represent HR, SR images, respectively. Considering the AE pretraining, ψ_{AE} is a differential approximation of a fidelity bias estimator. Thus, $\mathcal{L}_{\text{AESOP}}$ is now fundamentally identical to only penalizing the SE factor of Eq.(2) or Eq.(7). Since these features are decoupled from the perceptual variance factors by construction, $\mathcal{L}_{\text{AESOP}}$ leads to increased fidelity without forcing visually important textures to vanish. Also, note that *auto-encoded* indicates the space after the decoder, not the bottleneck.

Final objective function. Since we focus on improving the *fidelity* loss term of the framework, we substitute $\mathcal{L}_{\text{pix}}$ with $\mathcal{L}_{\text{AESOP}}$, leading to the overall objective function as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{AESOP}}\mathcal{L}_{\text{AESOP}} + \lambda_2\mathcal{L}_{\text{percep}} + \lambda_3\mathcal{L}_{\text{artif}} + \lambda_4\mathcal{L}_{\text{adv}}, \quad (9)$$

where λ_{AESOP}, λ_2, λ_3, λ_4 are coefficients for each loss factors. The overall pipeline of our training strategy is visualized in Fig.2. Based on the constructions above, the proposed $\mathcal{L}_{\text{AESOP}}$ provides reconstruction guidance without facing conflicts with the perceptual-quality-oriented losses. This indicates that both the $\mathcal{L}_{\text{AESOP}}$ and perceptual-quality-oriented losses can converge to an optimal point, leading to increased performance in terms of perception-distortion trade-off [2]. Accordingly, while typical methods multiply a very small coefficient to the reconstruction loss (generally chosen as 0.01 [37, 54]) to prevent blurring effects, we let $\lambda_{\text{AESOP}} = 1$. This way, we can provide significantly stronger reconstruction guidance without suffering from unintended blurring, leading to both lower levels of artifacts [37] and enhanced realism. For the other coefficients, we follow our baseline [37] settings and choose $\lambda_2 = 1$, $\lambda_3 = 1$, $\lambda_4 = 0.005$.

AE collapse. Eq.(8) leads to a trivial solution when the AE outputs the same value regardless of the input. To prevent this, we keep the AE frozen when training the SR network.

	Backbone	RRDB					SwinIR			
Metrics	Benchmark	ESRGAN	SPSR	LDL	AESOP	AESOP†	+GAN	LDL	AESOP*	AESOP
Recon. Objective		$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\textbf{AESOP}}$	$\mathcal{L}_{\textbf{AESOP}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\text{pix}}$	$\mathcal{L}_{\textbf{AESOP}}$	$\mathcal{L}_{\textbf{AESOP}}$
Patch Size (Training)		128	128	128	128	256	256	256	256	256
LPIPS ↓	Set14	0.1241	0.1207	0.1132	**0.1067**	0.1053	0.1160	0.1091	**0.1023**	0.1027
	Manga109	0.0649	0.0672	0.0544	**0.0525**	0.0494	0.0542	0.0469	**0.0440**	0.0461
	General100	0.0879	0.0862	0.0796	**0.0784**	0.0734	0.0796	0.0740	**0.0717**	0.0710
	Urban100	0.1229	0.1184	0.1084	**0.1064**	0.1033	0.1077	0.1021	**0.0961**	0.0945
	DIV2K-val	0.1154	0.1099	0.0999	**0.0977**	0.0936	0.1038	0.0944	**0.0909**	0.0893
	BSD100	0.1616	0.1609	0.1535	**0.1515**	0.1443	-	0.1572	**0.1441**	0.1385
	LSDIR	0.1378	0.1312	0.1180	**0.1152**	0.1123	-	0.1132	**0.1094**	0.1071
DISTS ↓	Set14	0.0951	0.0920	0.0866	**0.0852**	0.0825	0.0930	0.0869	**0.0809**	0.0819
	Manga109	0.0471	0.0463	**0.0355**	0.0360	0.0356	0.0365	**0.0315**	0.0327	0.0328
	General100	0.0874	0.0884	0.0801	**0.0798**	0.0773	0.0835	0.0794	**0.0768**	0.0762
	Urban100	0.0880	0.0849	**0.0793**	**0.0793**	0.0768	0.0835	0.0800	**0.0751**	0.0742
	DIV2K-val	0.0593	0.0546	0.0526	**0.0518**	0.0484	0.0531	0.0507	**0.0469**	0.0459
	BSD100	0.1165	0.1176	0.1163	**0.1117**	0.1089	-	0.1185	**0.1078**	0.1072
	LSDIR	0.0764	0.0699	0.0650	**0.0641**	0.0612	-	0.0650	**0.0601**	0.0591
PSNR ↑	Set14	26.594	26.860	27.228	**27.361**	27.246	27.282	27.526	**27.822**	27.421
	Manga109	28.413	28.561	29.620	**29.973**	29.747	29.345	30.143	**30.453**	30.061
	General100	29.425	29.424	30.289	**30.482**	30.251	30.104	30.441	**30.752**	30.401
	Urban100	24.365	24.804	25.459	**25.630**	25.541	25.736	26.231	**26.398**	26.148
	DIV2K-val	28.175	28.182	28.819	**29.079**	28.910	28.784	29.117	**29.543**	29.137
	BSD100	25.313	25.501	25.954	**26.080**	25.904	-	26.216	**26.405**	25.930
	LSDIR	23.882	24.232	24.663	**24.933**	24.845	-	25.129	**25.419**	25.038
SSIM ↑	Set14	0.7144	0.7254	0.7358	**0.7402**	0.7371	0.7407	0.7478	**0.7578**	0.7438
	Manga109	0.8595	0.8590	0.8734	**0.8827**	0.8802	0.8796	0.8880	**0.8949**	0.8880
	General100	0.8095	0.8091	0.8280	**0.8335**	0.8269	0.8305	0.8347	**0.8415**	0.8328
	Urban100	0.7341	0.7474	0.7661	**0.7724**	0.7697	0.7786	0.7918	**0.7947**	0.7884
	DIV2K-val	0.7759	0.7720	0.7897	**0.7978**	0.7951	0.7911	0.8011	**0.8121**	0.8023
	BSD100	0.6527	0.6596	0.6813	**0.6841**	0.6783	-	0.6923	**0.6982**	0.6813
	LSDIR	0.6866	0.6966	0.7117	**0.7220**	0.7202	-	0.7316	**0.7397**	0.7289

Table 1. Quantitative comparison between AESOP (Ours) and baseline methods. The best results of each group are highlighted in **bold**. AESOP* indicates only training 200K iterations, AESOP† indicates training with a larger patch.

5. Experiments

5.1. Benchmark evaluation

Experimental setup. We employ benchmark datasets including Set14 [62], Manga109 [44], General100 [14], Urban100 [19], DIV2K [1], BSD100 [43], LSDIR [35]. For both the AE pretraining and the SR network training, we use DF2K, a combination of DIV2K [1] and Flickr2K [38]. We report PSNR and SSIM [57] scores for distortion metrics and LPIPS [66], DISTS [12] for perceptual quality metrics. We use both RRDB-based models: ESRGAN [54], SPSR [42], LDL [37], CALGAN[48]; and SwinIR-based versions of each, if available. AESOP is used interchangeably to indicate either $\mathcal{L}_{\text{AESOP}}$ or the SR networks trained with Eq.(9). Refer to the Appendix for details on experimental settings.

Quantitative comparison. In Tab.1, we perform a quantitative evaluation against baseline GAN-based perceptual SR methods. Against all RRDB [54] based baseline methods, AESOP significantly improves both distortion metrics and perceptual scores. We find the key factor of this improvement as the carefully designed feature space and pretrained AE. It provides an increased level of reconstruction guidance, specifically to fidelity bias factor, while providing additional freedom to perceptual-quality-oriented losses. For

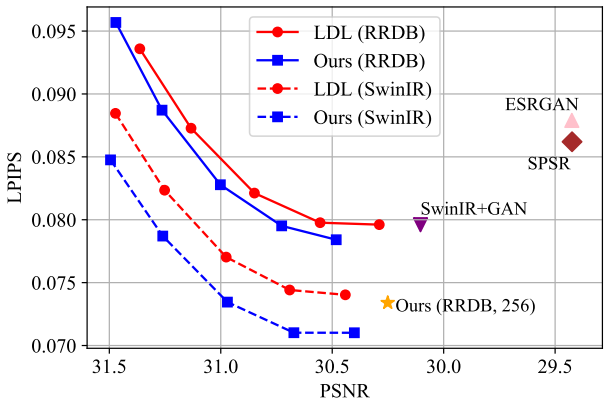
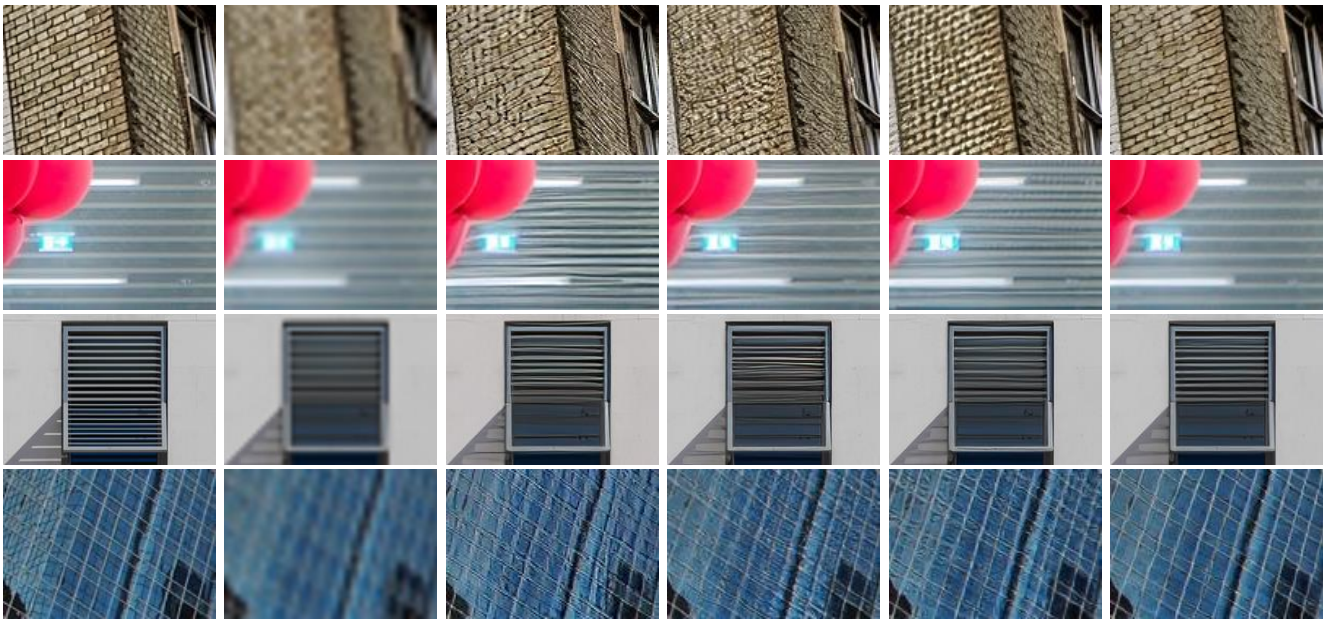


Figure 4. The PD trade-off curve. The backbone and training patch size are indicated; if not specified, the default patch size is 128.

SwinIR [36] backbone methods, we report both the final results at 300K iterations and intermediate training results at 200K iterations (indicated as AESOP*). At 200K, it can be seen that AESOP leads to improvements in both fidelity scores and perceptual metrics, similar to the RRDB-backbone. When we fully train our model up to 300K, we observe further enhancements in perception scores. Since improved perception leads to lower fidelity due to the PD trade-off, we provide the PD trade-off curves in Fig.4, with



(a) HR (b) Bicubic (c) ESRGAN (d) SPSR (e) LDL (f) AESOP (Ours)

Figure 5. Visual comparison of AESOP with baseline methods for the $\times 4$ SR task on the RRDB backbone. Our method produces images with fewer visual artifacts. See the Appendix for more visual examples of AESOP's improvement in realism and fine details.

PSNR and LPIPS scores on the General100 dataset. AESOP leads to improved trade-off relationships on both RRDB and SwinIR backbones. Additionally, we provide RRDB-backbone results using a training HR patch size of 256 (denoted as AESOP†) to align with the training settings of SwinIR. AESOP also improves realism for real-world SR tasks as in Tab.2. Refer to Appendix.B for more results.

Qualitative comparison. In Fig.5, baseline methods often suffer from unpleasant GAN artifacts [37] while AESOP presents a significantly lower level of these artifacts. This is due to the strong reconstruction guidance $\mathcal{L}_{\text{AESOP}}$ provides, since it does not require a small scaling factor. Additionally, refer to the Appendix for visual examples of AESOP improving realism on fine details and complex textures.

Ablation studies. To verify the effects of each component, we perform ablation studies on AESOP (128). In Tab.3, we report DISTS and PSNR scores for each setting on DIV2K validation set. The first line indicates the baseline setting that matches LDL, where none of our proposed components are applied. Tab.3.(a) indicates employing a decoder, but using a simple bicubic downsampling operation instead of the encoder. As reported, the perceptual quality and fidelity are improved even when solely utilizing the decoder due to its ability in offering stronger reconstruction loss without conflict with perceptual objectives. However, the usage of bicubic downsampling with a decoder corresponds to the bottleneck collapse, where no loss backpropagates if the SR image downsamples to the LR image, thus, leading to slightly lower performance against our full method. In Tab.3.(b), we further introduce a learnable encoder together with the decoder, but without Eq.(4) (i.e., without estimating the LR image). Given that both fidelity bias factors and perceptual variance factors are determined by the LR counterpart, the model cannot properly estimate the fidelity bias and the perceptual variance. Consequently, the perceptual variance factor can be penalized, which is not intended, leading to the lowest perceptual score. Our full model in Tab.3.(c) further employs Eq.(4). Thus, it better models the LR than Tab.3.(b), while also avoiding the bottleneck collapse which Tab.3.(a) has suffered, leading to the best scores in terms of perceptual quality.

Dataset	Method	Recon. Obj.	NIQE↓	MANIQA↑
RealSRv3 [3]	Real-ESRGAN	$\mathcal{L}_{\text{pix}}$	4.6790	0.3662
	AESOP (Ours)	$\mathcal{L}_{\text{AESOP}}$	**4.2337**	**0.4136**
DRealSR [58]	Real-ESRGAN	$\mathcal{L}_{\text{pix}}$	4.7152	0.3404
	AESOP (Ours)	$\mathcal{L}_{\text{AESOP}}$	**4.1922**	**0.3917**

Table 2. Quantitative results of AESOP in real-world settings. Refer to the Appendix for further results, including visual examples.

	Decoder	Encoder	$\mathcal{L}_{\text{LR}}^{\text{rec}}$	DISTS ↓	PSNR ↑
Baseline (LDL)				0.0526	28.819
Config-(a)	✓			0.0521	29.060
Config-(b)	✓	✓		0.0526	**29.150**
Config-(c) (Ours)	✓	✓	✓	**0.0518**	29.079

Table 3. Ablation study on each component of AESOP (128).

5.2. Analysis

Spectral analysis. Several prior works [11, 69] employ low-pass filtering (LPF) to avoid the blurring effect of $\mathcal{L}_{\text{pix}}$. To identify the difference between our AE and LPF, we analyze the spectral magnitudes of Fig.6.(a) after applying AE

and LPF, in Fig.6.(c)-(d), respectively. As can be seen, LPF blindly removes all HF components, contrary to AE, where certain patterns in the HF regions are preserved. As discussed in Sec. 3, the remaining HF components in Fig.6.(c) are factors that can be learned by pixel-level regression, and the removed HF components are the non-regressable factors (VE) that lead to blurring when minimized.

Here, we focus on the *remaining* HF components and visualize it in Fig.6.(e), by taking the absolute difference between Fig.6.(c)-(d). Simple object edges are highlighted, which are regions where even fidelity-oriented SR networks (trained with pixel-level regression loss, $\mathcal{L}_{\text{pix}}$), can also sharply reconstruct. This indicates that specific HF components *can* be regressed, and importantly, these HF components *cannot* be disentangled from other HF components by band-pass filters (e.g., LPF) or frequency selection [7, 8]; since they are intertwined within the same frequency band.

Meanwhile, LPF is expected to provide limited supervision in these regressable HF components, leading to degraded performance. To validate this statement, we compare AESOP against LPF in Tab.5.(b). We apply LPF on SR and HR images before calculating $\mathcal{L}_{\text{pix}}$ on top of our baseline method LDL. A notable degradation in performance can be observed, highlighting the superiority of our AE-based method in maintaining high-quality reconstruction guidance over conventional frequency-based methods.

Overall, we conclude that our AE can successfully disentangle fidelity biases and perceptual variances, by capturing specific HF components that 1) cannot be obtained by simple frequency selection, 2) but can be learned via regression, and 3) significantly contribute to improved fidelity.

Loss map comparison. Fig.3 visualizes key components of $\mathcal{L}_{\text{AESOP}}$ and $\mathcal{L}_{\text{pix}}$. In Fig.3.(e), $\mathcal{L}_{\text{pix}}$ cannot distinguish perceptual variance factors and fidelity bias factors. Thus, visually important fine-textures are penalized, leading to a blurry result as in Fig.3.(b). Meanwhile, $\mathcal{L}_{\text{AESOP}}$ in Fig.3.(f) successfully extracts and penalizes only the fidelity bias factor Fig.3.(d), leading to increased realism as in Fig.3.(c).

Fidelity bias estimation. Since $\mathcal{L}_{\text{AESOP}}$ does not lead to blurring, we do not multiply a small scaling factor. Accordingly, we can provide significantly stronger guidance on fidelity biases. Here, we measure how well each network estimates the fidelity biases, apart from PSNR scores which are influenced by perceptual variances. To do this, we introduce *AE-PSNR*, which measures the PSNR between auto-encoded SR and HR images. This score reflects how well an image captures the fidelity bias of the reference image. However, since AESOP is trained using the AE, there may be unintended biases introduced by the AE itself. Thus, we additionally report LR-PSNR as an unbiased metric independent of the AE, which measures the PSNR between downscaled SR and the original LR images. This captures

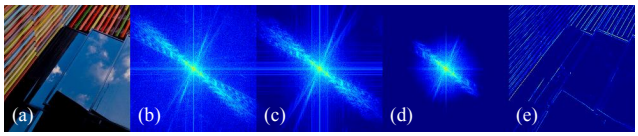


Figure 6. Visual comparison between AE and LPF. **(a)** Original image. **(b)** Original image in spectral domain. **(c)** Forwarding through AE. **(d)** Applying LPF. **(e)** Absolute difference between (c) and (d). (Electronic viewer highly recommended.)

	Method	Set14	Mg109	Gen100	Urb100	DIV2K	B100	LSDIR
AE-	LDL [37]	31.525	33.215	33.994	29.374	32.855	29.792	29.071
	Ours	**32.111**	**33.635**	**34.535**	**29.666**	**33.490**	**30.366**	**29.552**
LR-	LDL [37]	46.899	49.135	48.663	47.404	48.084	45.494	45.731
	Ours	**48.245**	**50.042**	**49.733**	**48.564**	**49.856**	**47.578**	**47.476**

Table 4. AE-PSNR and LR-PSNR scores with SwinIR-backbone.

	Manga109	General100	Urban100	DIV2K100
Ours	**29.97/.0525**	**30.48/.0784**	**25.63**/.1064	**29.08/.0977**
(a)	29.78/.0534	30.35/.0789	25.55/**.1054**	28.97/.0982
LDL	29.62/.0544	30.29/.0796	25.46/.1084	28.82/.0999
(b)	29.55/.0545	30.20/.0801	25.39/.1090	28.75/.1005

Table 5. PSNR/LPIPS scores. **(a)** AESOP with SRResNet-based AE (ψ). **(b)** Applying LPF before calculating $\mathcal{L}_{\text{pix}}$ with LDL [37].

how well the fidelity biases align, but without being influenced by the AE. However, note that this measure only reflects the LF feature, a subcomponent of the fidelity bias. In Tab.4, AESOP shows improvements on both AE-PSNR and LR-PSNR scores, demonstrating the superiority of $\mathcal{L}_{\text{AESOP}}$ against scaled $\mathcal{L}_{\text{pix}}$ in effectively reducing the SE term. See the Appendix for scores on RRDB-backbone methods.

Architectural choice of AE. Performance of AESOP with SRResNet-based [30] AE is reported in Tab.5.(a). We observe a slight performance drop against our RRDB-based AE, but it is still superior against LDL. While this indicates that AESOP relies on a well-performing AE, this reliance does not pose a practical issue. In our pretraining framework, we initialize the decoder of AE as the fidelity-oriented SR network, which is expected to be already in place under the SRGAN-based training framework.

6. Conclusion

This work analyzes limitations of $\mathcal{L}_{\text{pix}}$ (i.e., the conventional pixel-level $\mathcal{L}_{\text{p}}$) in the context of perceptual SR. Further, we highlight the shortcomings of prior circumvention to avoid blurring, in terms of fidelity biases and perceptual variance factors. We tackle this issue by introducing $\mathcal{L}_{\text{AESOP}}$, a novel reconstruction loss that separates fidelity bias factors from perceptual variance factors using an AE, pretrained for a reconstruction task. This allows us to focus on enhancing fidelity while preserving the visual quality of SR images. Experimental results validate that the proposed method leads to significant improvement in the perceptual SR task.

Acknowledgement This work was supported in part by MSIT/IITP (No. 2022-0-00680, 2020-0-01821, RS-2019-II190421, RS-2024-00459618, RS-2024-00360227, RS-2024-00437102, RS-2024-00437633), and MSIT/NRF (No. RS-2024-00357729).

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 6

[2] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 1, 3, 5, 6

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019. 7, 2

[4] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. In *European Conference on Computer Vision*, pages 170–188. Springer, 2022. 2

[5] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 2

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 2

[7] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 8

[8] Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, and Alois Knoll. Selective frequency network for image restoration. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[9] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 2

[10] Mauricio Delbracio, Hossein Talebei, and Pevman Milanfar. Projected distribution loss for image enhancement. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2021. 3

[11] Xin Deng, Ren Yang, Mai Xu, and Pier Luigi Dragotti. Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3076–3085, 2019. 1, 3, 7

[12] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 6

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

[14] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016. 6

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[16] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[18] Chih-Chung Hsu, Chia-Ming Lee, and Yi-Shiuan Chou. Drct: Saving image super-resolution away from information bottleneck. *arXiv preprint arXiv:2404.00722*, 2024. 2, 4, 11, 12

[19] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 6

[20] Sangeek Hyun and Jae-Pil Heo. Varsr: Variational super-resolution network for very low resolution images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, pages 431–447. Springer, 2020. 2

[21] Gareth M James. Variance and bias for general loss functions. *Machine learning*, 51:115–135, 2003. 3

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1, 2, 4

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[24] 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*, pages 5148–5157, 2021. 2

[25] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[27] Idan Kligvasser and Tomer Michaeli. Sparsity aware normalization for gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8181–8190, 2021. 3

[28] Idan Kligvasser, Tamar Shaham, Yuval Bahat, and Tomer Michaeli. Deep self-dissimilarities as powerful visual fingerprints. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[29] Denis Kuznedelev, Valerii Startsev, Daniil Shlenskii, and Sergey Kastryulin. Does diffusion beat gan in image super resolution? *arXiv preprint arXiv:2405.17261*, 2024. 3

[30] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 2, 3, 8

[31] MinKyu Lee and Jae-Pil Heo. Noise-free optimization in early training steps for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2920–2928, 2024. 1, 3

[32] Soochan Lee, Junsoo Ha, and Gunhee Kim. Harmonizing maximum likelihood with gans for multimodal conditional generation. *arXiv preprint arXiv:1902.09225*, 2019. 3

[33] Bingchen Li, Xin Li, Hanxin Zhu, Yeying Jin, Ruoyu Feng, Zhizheng Zhang, and Zhibo Chen. Sed: Semantic-aware discriminator for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25784–25795, 2024. 3

[34] Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer-based image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 2

[35] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023. 6

[36] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 2, 6, 1, 13, 14

[37] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022. 3, 4, 5, 6, 7, 8, 1, 2

[38] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 6

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

[40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[41] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 715–732. Springer, 2020. 2

[42] Cheng Ma, Yongming Rao, Jiwen Lu, and Jie Zhou. Structure-preserving image super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7898–7911, 2021. 2, 3, 6

[43] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 416–423. IEEE, 2001. 6

[44] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017. 6, 7

[45] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021. 2

[46] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2

[47] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 191–207. Springer, 2020. 2

[48] JoonKyu Park, Sanghyun Son, and Kyoung Mu Lee. Content-aware local gan for photo-realistic super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10585–10594, 2023. 3, 6, 2, 4, 7, 15

[49] Seung Ho Park, Young Su Moon, and Nam Ik Cho. Perception-oriented single image super-resolution using optimal objective estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1725–1735, 2023. 3

[50] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. 3

[51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[52] 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*, pages 2555–2563, 2023. 2

[53] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 3

[54] 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 *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1, 3, 4, 5, 6, 2, 9, 10, 15

[55] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 1, 2

[56] Xintao Wang, Liangbin Xie, Ke Yu, Kelvin C.K. Chan, Chen Change Loy, and Chao Dong. BasicSR: Open source image and video restoration toolbox. https://github.com/XPixelGroup/BasicSR, 2022. 1

[57] 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. 6

[58] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 101–117. Springer, 2020. 7, 2

[59] Liangbin Xie, Xintao Wang, Xiangyu Chen, Gen Li, Ying Shan, Jiantao Zhou, and Chao Dong. Desra: Detect and delete the artifacts of gan-based real-world super-resolution models. 2023. 3, 4

[60] 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*, pages 1191–1200, 2022. 2

[61] 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*, pages 25669–25680, 2024. 3

[62] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010. 6

[63] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022. 2

[64] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[65] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 1, 2

[66] 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*, pages 586–595, 2018. 6, 1

[67] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 2

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

[69] Yuehan Zhang, Bo Ji, Jia Hao, and Angela Yao. Perception-distortion balanced admm optimization for single-image super-resolution. In *European Conference on Computer Vision*, pages 108–125. Springer, 2022. 1, 3, 7

[70] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. *Advances in neural information processing systems*, 33:3499–3509, 2020. 2