Training Generative Image Super-Resolution Models by Wavelet-Domain Losses Enables Better Control of Artifacts*

Cansu Korkmaz, A. Murat Tekalp, Zafer Dogan
College of Engineering and KUIS AI Center, Koc University
https://github.com/mandalinadagi/WGSR

Abstract

Super-resolution (SR) is an ill-posed inverse problem, where the size of the set of feasible solutions that are consistent with a given low-resolution image is very large. Many algorithms have been proposed to find a "good" solution among the feasible solutions that strike a balance between fidelity and perceptual quality. Unfortunately, all known methods generate artifacts and hallucinations while trying to reconstruct high-frequency (HF) image details. A fundamental question is: Can a model learn to distinguish genuine image details from artifacts? Although some recent works focused on the differentiation of details and artifacts, this is a very challenging problem and a satisfactory solution is yet to be found. This paper shows that the characterization of genuine HF details versus artifacts can be better learned by training GAN-based SR models using wavelet-domain loss functions compared to RGB-domain or Fourier-space losses. Although wavelet-domain losses have been used in the literature before, they have not been used in the context of the SR task. More specifically, we train the discriminator only on the HF wavelet sub-bands instead of on RGB images and the generator is trained by a fidelity loss over wavelet subbands to make it sensitive to the scale and orientation of structures. Extensive experimental results demonstrate that our model achieves better perception-distortion trade-off according to multiple objective measures and visual evaluations.

1. Introduction

Single image super-resolution (SR) aims to reconstruct high-frequency (HF) details missing in low-resolution (LR) images. Early deep-learning based SR works employed simple convolutional neural networks (CNN), trained by pixel-wise l_1 and l_2 fidelity losses [11, 26]. They were followed by better models, which adopted residual [29, 32, 69] and dense connections [58, 70]. Later, the spatial attention, channel attention [7, 43, 46, 64, 69] and transformer net-

* This work is supported by TUBITAK 2247-A Award No. 120C156, TUBITAK 2232 Int. Fellowship for Outstanding Researchers Award No. 118C337, KUIS AI Center, and Turkish Academy of Sciences (TUBA).

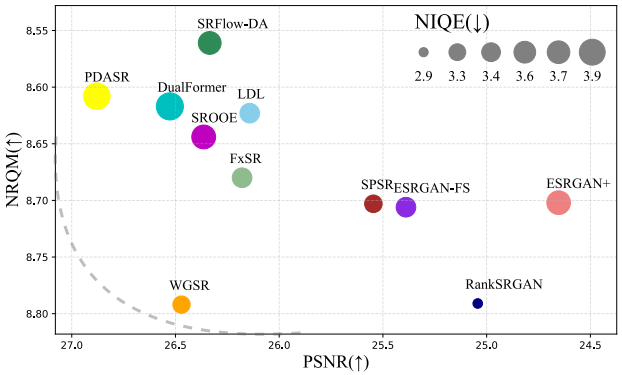

Figure 1. Perception-distortion trade-off performance of our model WGSR vs. state-of-the-art methods on the PSNR-NRQM plane. Dashed curve shows the theoretical limit explained in [3].

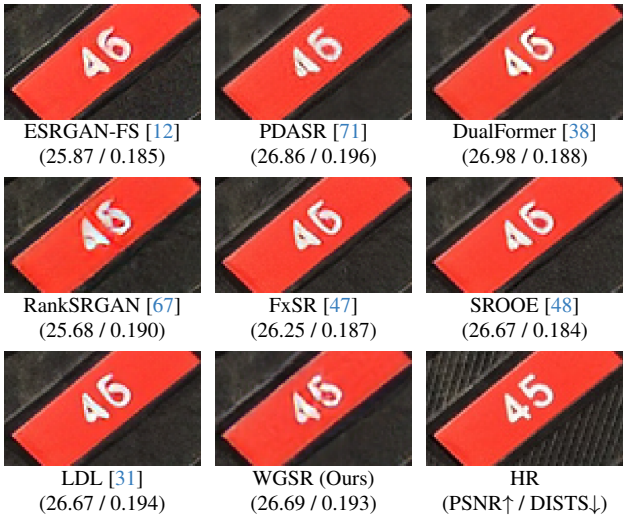

Figure 2. Visual performance of recent ×4 SR methods on a crop from Urban100 dataset (img-6) [20]. SOTA methods reconstruct "5" as "6", whereas the opening in the lower part of "5" is visible in our result confirming that our model strikes a better balance between fidelity and visual quality. Note PSNR, DISTS and other quantitative scores are not good indicators of such artifacts.

works [30, 68] have demonstrated impressive performance in terms of peak-signal-to-noise ratio (PSNR) and structural similarity measure (SSIM). However, minimization of mean-square error favors a probability-weighted average of all feasible SR outputs; hence, models that are optimized based only on fidelity losses produce overly smoothed images that lack HF details.

In order to generate visually more appealing results, generative SR models such as generative adversarial networks (GANs) [29, 40, 52, 60, 67][10, 24, 29, 39, 65], flow models [27, 35], and diffusion models [15, 49, 51, 53] have been proposed. Generative SR models aim to sample predicted SR images from a distribution that is similar to that of ground-truth (GT) images. However, they are known to hallucinate HF details and produce structural artifacts. Flow and diffusion models perform stochastic sampling in the sense that a single model can generate many samples. Hence, they allow less control per sample on learning details vs. artifacts. In this paper, we focus on conditional GAN-SR models, where a single trained model generates a single SR image sample. GAN models are trained by a weighted sum of pixel-wise fidelity and adversarial (discriminator) losses [16]. Additional perceptual losses, such as the VGG loss [29], the texture matching loss [57], and the content loss [42] have been suggested to enforce feature-level similarity between SR and GT images to alleviate hallucinations and artifacts. However, perceptual losses are not sufficiently effective to control hallucinations and artifacts.

The perception-distortion (PD) trade-off hypothesis [3] states there is a bound beyond which any perceptual quality improvement (measured by a no-reference metric) comes at the expense of increased distortion (measured by a full-reference metric). Finding the best trade-off between fidelity and perceptual quality is not a well-defined optimization problem mainly because no quantitative perceptual image quality measure correlates well with human preferences. Recognizing this, recent SR challenges require consistency of SR reconstructions with the LR observations under the forward degradation model (also called feasible solutions) and conduct human evaluations for visual quality [17, 36, 37]. Yet, the size of the set of feasible solutions is very large, and determining which feasible solutions contain genuine image details and which contain artifacts or hallucinations is extremely challenging even for humans.

In this paper, we propose a novel GAN-SR framework that uses wavelet-domain losses to suppress hallucinations and artifacts for a better PD trade-off. We define fidelity and adversarial losses over the subbands of the stationary wavelet transform (SWT), where the scale and orientation of decomposed image features are well represented. Since the SWT decomposes an image without sub-sampling, it is able to provide the distinctive local features of low-frequency (LF) and HF subbands. Enforcing the reconstructed SR images to preserve local statistics within different subbands of HR images as an optimization goal enables the model to learn image details with different scales and orientations for a better PD trade-off.

Our wavelet-guided super-resolution (WGSR) model provides a better PD trade-off in the NRQM vs. PSNR plane compared to other state-of-the-art (SOTA) methods as shown in Fig. 1, where our NRQM score is the best among other methods with similar PSNR and our PSNR score is higher than RankSRGAN, which has similar NRQM score. Also, Fig. 2 demonstrates a visual comparison of our method and other SOTA methods. WGSR, shows remarkable performance by regulating easily visible artifacts, e.g., the opening in the lower part of "5" is visible in our result, while other SOTA methods reconstruct "5" as "6." Note that quantitative scores, such as PSNR, DISTS and others, are not good indicators of such artifacts. To the best of our knowledge, our method is the first adversarial training scheme that employs wavelet guidance for artifact control, which can be applied to any GAN-SR model. To summarize, our primary contributions are:

- We propose a wavelet-domain fidelity loss (a weighted combination of l_1 losses on different wavelet sub-bands instead of the conventional RGB-domain l_1 loss), which is sensitive to the scale and orientation of local structures in images better observed in the SWT subbands.
- We propose utilizing an SWT-domain discriminator for adversarial training in order to control HF artifacts. We show that training the discriminator over HF wavelet subbands allows better control of the optimization landscape to segregate artifacts from genuine image details compared to the traditional RGB-domain discriminator.
- We show that combining our proposed wavelet-guided training scheme with the RGB-domain DISTS perceptual loss (instead of the conventional VGG-based LPIPS loss) significantly improves fidelity (up to 0.5 dB in PSNR) with minimal (less than 1%) loss in perceptual quality.

2. Related Work

GAN-based SR. GANs [16] offer a principled approach to achieve PD trade-off by controlling the weights of perceptual and fidelity losses in order to generate realistic images. To improve perceptual quality, Johnson *et al.* [24] proposed a perceptual loss. Ledig *et al.* [29] proposed SRGAN with adversarial training along with the SRResNet generator. Wang *et al.* [60] proposed the ESRGAN with the Residual-in-Residual Dense Block (RRDB) architecture which has been employed as a standard backbone in many SOTA GAN-SR methods. Later, Rakotonirina *et al.* [50] improved ESRGAN by additional noise injection and proposed ESRGAN+. Zhang *et al.* [67] presented a Ranker that learns the behavior of perceptual metrics in RankSRGAN. Ma *et al.* [40] proposed SPSR attenuate geometric

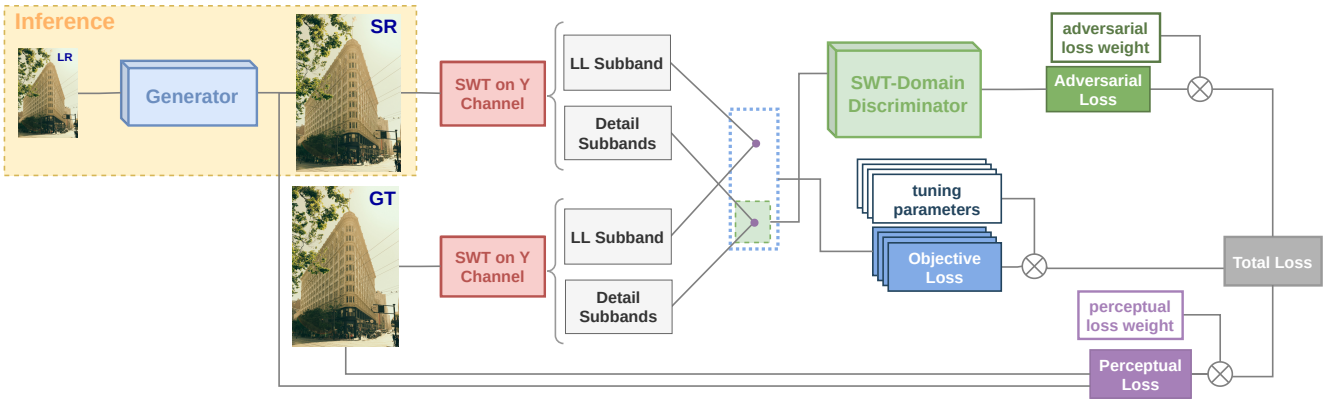

Figure 3. Overview of the proposed GAN-SR framework guided by wavelet-domain losses, where the strength of the adversarial loss is tuned for each subband to control artifacts and the discriminator learns to decide whether the generated detail subbands are real or fake.

distortions by preserving structure. Liang *et al.* [31] suggested a locally discriminative learning framework LDL by externally computing a probability map of each pixel being artifacts based on patch-level residual variances. Park *et al.* [47] introduced Flexible Style Image Super-Resolution (FxSR), which optimizes SR network with image-specific objectives without considering the regional characteristics. Later, in SROOE, Park *et al.* [48] proposed optimal objective estimation depending on perceptual and objective image maps. These methods [31, 47, 48] coexist with the computational burden of a large number of image maps. On the contrary, our wavelet-loss guided model does not require explicit calculation of an artifact map and inherently learns to suppress artifacts while retaining genuine details.

Training GANs by Frequency Domain Losses. Many studies have proposed frequency-related losses that better control the PD trade-off and ease the training of GANs [6, 12–14, 55, 72]. Fritsche *et al.* [12] proposed ESRGAN-FS, and the adversarial loss computed solely on the high-pass filtered images. Zhou *et al.* [72] introduced CARB GAN-FS with two discriminator models that treat LF and HF components separately. In [22], Jiang *et al.* proposed a focal frequency loss to alleviate the generation of frequency components that are hard to synthesize by means of a weighted Fourier space distance. Fuoli *et al.* [13] suggested Fourier domain discriminator to eliminate spectral discrepancies. Recently, Luo *et al.* [38] introduced DualFormer, which utilizes spatial and spectral discriminators simultaneously. However, DFT domain (spectral) losses cannot localize HF image features according to scale and orientation, unlike wavelet decompositions, to characterize genuine details vs. artifacts.

Modeling SR in the Wavelet Domain. Wavelet decomposition based approaches played crucial role in various computer vision tasks including GAN-inversion [33, 45], generative modeling [14, 49, 56], face-aging [5, 34], video compression [59], medical and thermal imaging [61, 66]. Wavelet-domain learning methods have also been applied to SR tasks [18, 19, 54, 62, 71]; but existing methods directly predict wavelet coefficients of SR images. Specifically, Deng *et al.* [8] proposed fusing images generated by objective and perceptual quality criteria via style transfer in the pixel domain. Later, Deng *et al.* [9] employed Wavelet Domain Style Transfer (WDST), which performs style transfer on wavelet subbands. Zhang *et al.* [71] proposed PDASR to achieve PD trade-off by a two-stage SR framework that employs a low-frequency content constraint. PDASR reconstructs different frequency subbands independently, which causes inconsistency between subbands and results in unnatural artifacts. In contrast, we are the first to train pixel-domain ESRGAN [60] model using weighted wavelet subband losses departing from conventional RGB $\ell 1$ loss. Our method, WGSR, is superior to others because predicting RGB pixels is easier than predicting sparse wavelet coefficients of detail bands, while unequal weighting of losses in different wavelet subbands enables learning structures with different scales and orientations. To the best of our knowledge, WGSR is the first GAN-based RGB-domain SR model guided by wavelet-domain losses.

3. WGSR: Wavelet-Guided SR Framework

We propose a novel adversarial training framework presented in Fig. 3 for GAN-SR models that suppresses HF hallucinations and artifacts to achieve better PD trade-off by (i) training the discriminator only on the HF subbands, (ii) introducing a wavelet domain distortion loss to guide the generator, and (iii) selecting more suitable perceptual loss that couples better with our optimization objective.

3.1. Rationale for using Wavelet-Domain Losses

The Stationary Wavelet Transform (SWT) allows multi-scale decomposition of images [21] into one LF subband

referred as LL and several HF (e.g., LH, HL, HH) subbands. The decomposition level of LL subband determines the number of HF subbands that convey the detailed information in horizontal, vertical, and diagonal directions, respectively. It is important to note that since the resolution is highly critical in SR tasks, we utilize the SWT rather than classical Discrete Wavelet Transform (DWT). The main difference of SWT is the removal of the decimation part in the DWT, hence, the SWT method inherently couples the scale/frequency information with spatial location.

The LL subband of the SWT decomposition has a significant effect on the fidelity of the reconstructed images [9]. Hence, it is crucial not to alter the existing frequencies or introduce new ones into the LL subband to attain low distortion. At the same time, the HF contents of an image that are aligned with the LL spatial contents need to be reconstructed to achieve photo-realistic images. To better demonstrate the key advantages of SWT-guided adversarial training, we apply 1-level SWT decomposition to HR image, to the result of ESRGAN+ [50] and to the result of our WGSR method and present these decompositions in Fig. 4. While training ESRGAN+ [50], there is no guidance provided by wavelet domain losses, hence it represents classical adversarial training approach, and the image generated by ESRGAN+ [50] contains exaggerated artifacts. When we closely examine the HF subbands, due to the orientation of structures in the image, the HL subband contains more hallucination with higher distortion, resulting to have the lowest PSNR score among other subbands. So, this specific patch of the ESRGAN+ [50] actually requires enhancement in the HL subband. However, extracting this information from the RGB image itself is much harder than doing it on the HL subband for the discriminator network and it fails to recognize this unnatural artifact in the vicinity of the window in the image. On the other hand, we optimize our discriminator network to separate the details from artifacts by only feeding the HF subbands as opposed to RGB images. As a result, our wavelet-guided optimization result shows significant improvement in all subbands, as well as in the final SR image; hence, it obtains remarkable photo-realistic result that contains genuine image details rather than hallucinated artifacts.

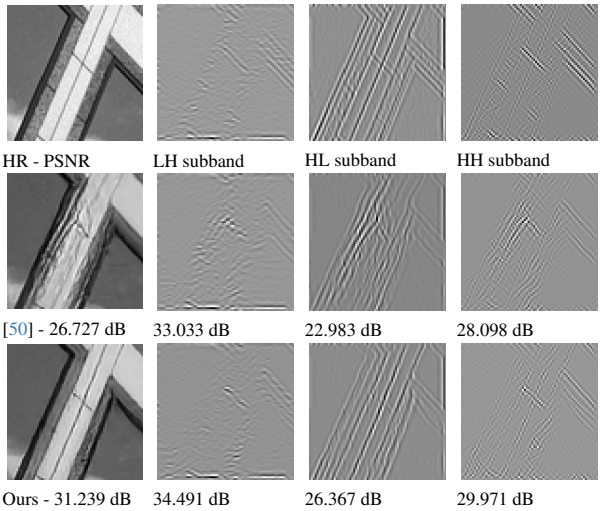

Figure 4. Illustration of our main premise that imposing different losses to different SWT subbands results in remarkable quantitative and qualitative performance improvements in GAN-based SR models. Specifically, enforcing fidelity loss on wavelet sub-bands instead of on RGB channels and running the discriminator only on detail (LH, HL, and HH) subbands helps eliminate visible artifacts caused by ESRGAN+ [50] and leads to better preservation of details. Overall scores of our method WGSR (PSNR: 26.33/DISTS: 0.115) outperform ESRGAN+ [50] (PSNR: 22.78/DISTS: 0.225).

3.2. Architecture

The proposed framework shown in Figure 3 consists of an RGB-domain generator and a SWT-domain discriminator, which are jointly trained using SWT-guided fidelity and adversarial, and pixel-domain perceptual losses. The framework is generic in the sense that any generator and any discriminator model can be easily plugged into this framework.

SWT-domain Discriminator We employ a discriminator network that is tasked to learn how "real" are the generated HF details (LH, HL, and HH subbands) compared to the ones that appear in the SWT decomposition of HR images. Our discriminator only evaluates the horizontal, vertical and diagonal details as opposed to evaluating RGB images, since they are crucial to control details vs. hallucinated artifacts. As shown in the last 3 columns of Fig. 4, LH, HL, and HH subbands convey sparse information, which simplifies the task of the discriminator and enables stable training. The training pipeline of the discriminator starts with YCbCr conversion of the generated image. The SWT decomposition is applied on the Y channel (Cb and Cr are discarded) to obtain LL, LH, HL and HH subbands. Only the details (LH,HL, HH) subbands are used to train the discriminator. The architecture of the discriminator consists of 9 convolution layers, whose kernel size alternates between 3x3 and 4x4, followed by 2D batch norms, and ReLU activation applied in between as in [25]. The number of output features of each convolutional layer increases from 64 to 512 and at the end, there are 2 linear layers with LeakyReLU activation which returns a 2D array to determine whether the HF subbands of the generated image resemble the ones of the GT image. Since this approach allows the discriminator to focus more on the relevant HF details of the generated images, which is where the artifacts are clearly separated from genuine details, it prevents hallucinations and eliminates distortions.

RGB-domain Generator The RRDB [60] architecture is selected as a backbone generator network, which consists

of 23 residual-in-residual dense blocks without batch norm. Except for the output layer, all convolutional layers use 3x3 kernels with 64 features, and Leaky ReLU is selected as the activation function. Since the generator network takes randomly cropped RGB patches during training we refer to it as RGB-domain generator. It is worth mentioning that our proposed training scheme with wavelet domain losses and SWT-domain discriminator can be coupled with any generator network architecture.

3.3. Training by SWT-Domain Losses

Instead of using the regular RGB-domain fidelity loss as in conventional GAN-SR methods, we define the SWT-domain fidelity loss, L_{SWT}, with corresponding tuning parameters. The flexibility of weighting the contribution of each subband individually enables adjusting the balance of fidelity and perceptual quality of the generated SR image. We sum the l_1 fidelity loss between SWT subbands of generated images x and the GT image y and average over a minibatch size denoted as $\mathbb{E}[.]$, given by

$$L_{SWT} = \mathbb{E}\big[\sum_j \lambda_j \big\| SWT(G(x))_j - SWT(y)_j \big\|_1\big] \quad (1)$$

where G denotes the generator model and λ_j are appropriate scaling factors to control the generated HF details to avoid hallucinations and disturbing visual artifacts appearing around fine-scale regular structures such as sharp lines/edges on windows, buildings, letters, or tree branches. When the lowest frequency (LL) subband contains flat regions or large-scale structures, it is important to preserve the shapes of objects to maintain the objective quality. So, we compute the adversarial loss term, given by equation 2, over the detail subbands (LH, HL, and HH) in order not to alter the existing frequencies or introduce new ones.

$$L_{adv,G} = -\mathbb{E}\big[log(1 - D(SWT(y)_*))\big] - \mathbb{E}\big[logD(SWT(G(x))_*)\big] \quad (2)$$

where D is the discriminator model, and $*$ indicates concatenation of details subbands.
Then, the overall loss for the generator is given by

$$L_G = L_{SWT} + \lambda_{adv} \cdot L_{adv,G} + \lambda_{perc} \cdot L_{perc} \quad (3)$$

where L_{perc} denotes the perceptual loss, measuring errors in the feature space provided by image quality assessment DISTS [10].
The loss term for the discriminator, which only takes the HF details subbands as input, is given by

$$L_D = -\mathbb{E}\big[logD(SWT(y)_*)\big] - \mathbb{E}\big[log(1 - D(SWT(G(x))_*))\big] \quad (4)$$

Since determination of the optimal values λ for each subband is not straightforward, we find the best PD trade-off point by searching empirically to be at $\lambda_{LL} = 0.1$, $\lambda_{LH} = \lambda_{HL} = 0.01$, $\lambda_{HH} = 0.05$, $\lambda_{adv} = 0.005$ and $\lambda_{perc} = 1$.

4. Experiments

4.1. Experimental Setup

Training Details. As a training set, we used 800 LR images from DIV2K [1] that are generated using the MATLAB bicubic downsampling kernel with a scaling factor of 4×. Randomly cropped 32×32 pixels of RGB LR patches on a minibatch of 16 are given to the generator. Then the loss terms are calculated after applying SWT to the Y-channel of generated images. The ADAM optimizer [28] with default settings $\beta_1 = 0.9$ $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ is selected for the optimization. We initialize training parameters of the generator with the pre-trained RRDB [60] weights and then perform 60k iterations with an initial learning rate of 10^{-4} which is halved after 50k iterations. Since wavelet loss is calculated during the training, it does not affect the runtime, hence the inference time of WGSR is the same as the inference time of RRDB [60].
Benchmarks and Metrics. To assess the generalization performance of our model, we report results on Set5 [2], Set14 [63], BSD100 [41], Urban100 [20] and DIV2K [1] validation dataset. We report PSNR and SSIM scores on the Y channel to demonstrate the objective quality of generated SR images. The perceptual quality of images is assessed via utilizing the full-reference metrics LPIPS [65], DISTS [10], and no-reference metrics NIQE [44], NRQM [39] and PI [4] on RGB images for a comprehensive evaluation. We also report LR-PSNR results on benchmarks to verify the LR-consistency of the predicted results. SR predictions must achieve at least 45 dB PSNR between the downsampled version of SR predictions and the corresponding LR images to satisfy the LR-Consistency criterion [36, 37].

4.2. Comparison with the state-of-the-art

Quantitative Comparison. Table 1 demonstrates quantitative comparison for ×4 SR methods and our proposed approach WGSR. We compare our method with the existing state-of-the-art methods including ESRGAN-FS [12], ESRGAN+ [50], SPSR [40], RankSRGAN [67], SRFlow-DA (heat=0.9) [23], LDL [31], FxSR (t=0.8) [47], PDASR [71] and SROOE (t=0.9)[48]. Our method WGSR improves the perceptual quality and the reconstruction accuracy simultaneously. Specifically, the table shows that our method yields the best perceptual scores in terms of NIQE, NRQM, and PI without significantly compromising objective quality. Also, in terms of distortion-oriented metrics such as PSNR and SSIM, our method provides better fidelity scores compared to other GAN-SR approaches. Our proposed network WGSR also exceeds 45 dB LR-PSNR which guarantees the original information conveyed by the LR images is preserved, thus, it does not suffer from LR-consistency, unlike ESRGAN-FS [12], ESRGAN+ [50], SPSR [40] and RankSRGAN [67]. To conclude, our pro-

Table 1. Quantitative comparison of the proposed wavelet decomposition-based optimization objective vs. other state-of-the-art methods for $\times 4$ SR task. The best and the second-best are marked in **bold** and underlined, respectively.

Benchmark	Metric	ESRGAN-FS [12]	ESRGAN+ [50]	SPSR [40]	RankSRGAN [67]	SRFlow-DA [23]	LDL [31]	FxSR [47]	PDASR [71]	DualFormer [38]	SROOE [48]	WGSR (1-lvl)	WGSR (2-lvl)
Dataset		DF2K	DIV2K	DIV2K	DIV2K	DF2K	DIV2K	DIV2K	DIV2K	DIV2K	DF2K	DIV2K	DIV2K
Set5	PSNR ↑	30.329	29.002	30.357	28.859	30.764	30.964	30.858	**31.728**	31.299	31.285	31.334	31.508
	SSIM ↑	0.844	0.801	0.843	0.823	0.855	0.860	0.855	**0.875**	0.869	0.867	0.866	0.869
	LPIPS ↓	0.065	0.100	0.065	0.077	0.084	0.066	**0.060**	0.078	0.066	0.064	0.065	0.068
	LPIPS-VGG ↓	0.172	0.224	0.167	0.194	0.189	0.151	0.163	0.178	0.149	**0.148**	0.168	0.162
	DISTS ↓	0.096	0.126	**0.092**	0.109	0.110	**0.092**	0.096	0.110	0.093	0.094	0.110	0.107
	NIQE ↓	4.320	4.710	4.215	**3.589**	5.600	4.602	4.642	5.520	5.225	5.067	4.175	4.270
	NRQM ↑	8.015	8.250	8.079	**8.613**	6.886	7.588	7.973	7.416	7.156	7.428	8.252	7.927
	PI ↓	3.385	3.380	3.337	**2.663**	4.439	3.681	3.565	4.286	4.210	3.884	3.160	3.239
	LR-PSNR ↑	42.950	42.860	43.630	38.210	49.940	46.590	50.210	**53.280**	42.929	51.020	49.911	50.868
Set14	PSNR ↑	26.415	25.923	26.564	25.797	27.123	27.096	27.115	**27.869**	27.394	27.278	27.395	26.689
	SSIM ↑	0.711	0.685	0.714	0.686	0.728	0.735	0.733	**0.751**	0.739	0.739	0.739	0.716
	LPIPS ↓	0.140	0.159	0.132	0.145	0.133	0.131	0.123	0.142	0.120	**0.116**	0.138	0.140
	LPIPS-VGG ↓	0.241	0.274	0.237	0.261	0.254	0.225	0.227	0.247	0.216	**0.215**	0.252	0.257
	DISTS ↓	0.102	0.126	0.098	0.112	0.113	0.098	0.097	0.112	0.092	**0.090**	0.112	0.110
	NIQE ↓	3.586	3.495	3.657	**3.220**	4.238	3.635	3.578	4.109	4.167	3.984	3.594	3.309
	NRQM ↑	8.029	8.046	8.056	**8.227**	7.843	7.907	7.992	7.818	7.821	7.928	7.930	8.196
	PI ↓	2.838	2.768	2.908	**2.519**	3.196	2.959	2.880	3.251	3.272	3.093	2.914	2.577
	LR-PSNR ↑	40.930	41.270	41.390	37.180	49.570	44.500	49.000	**50.510**	41.678	49.150	49.023	48.937
BSD100	PSNR ↑	25.389	24.653	25.546	25.043	26.335	26.142	26.179	**26.879**	26.527	26.364	26.471	26.372
	SSIM ↑	0.658	0.614	0.659	0.639	0.684	0.682	0.685	**0.703**	0.691	0.693	0.696	0.684
	LPIPS ↓	0.166	0.211	0.161	0.183	0.191	0.163	0.157	0.187	0.158	**0.153**	0.187	0.174
	LPIPS-VGG ↓	0.269	0.313	0.263	0.285	0.286	0.244	0.253	0.272	0.242	**0.241**	0.283	0.282
	DISTS ↓	0.119	0.151	0.118	0.129	0.145	0.118	0.118	0.136	0.119	**0.116**	0.137	0.132
	NIQE ↓	3.386	3.675	3.261	**2.903**	3.603	3.383	3.386	3.902	3.957	3.684	3.428	3.243
	NRQM ↑	8.706	8.702	8.703	8.791	8.561	8.623	8.680	8.608	8.617	8.644	8.792	**8.793**
	PI ↓	2.402	2.531	2.335	2.086	2.631	2.473	2.422	2.779	2.796	2.576	**2.053**	2.065
	LR-PSNR ↑	39.910	41.530	40.990	37.510	**49.920**	43.690	49.260	49.830	42.306	49.610	49.046	48.915
Urban100	PSNR ↑	24.556	23.235	24.795	24.121	25.632	25.491	25.668	**26.279**	25.686	25.939	25.779	25.606
	SSIM ↑	0.743	0.707	0.747	0.719	0.763	0.767	0.772	**0.785**	0.773	0.779	0.781	0.777
	LPIPS ↓	0.124	0.143	0.119	0.143	0.129	0.110	**0.109**	0.123	0.115	0.108	0.135	0.135
	LPIPS-VGG ↓	0.222	0.248	0.216	0.249	0.241	**0.197**	0.204	0.223	0.200	0.199	0.243	0.243
	DISTS ↓	0.090	0.104	0.085	0.106	0.115	**0.082**	0.087	0.102	0.085	0.085	0.108	0.101
	NIQE ↓	3.803	3.639	3.686	3.712	4.361	3.777	3.801	4.012	4.148	3.906	3.526	**3.326**
	NRQM ↑	6.652	6.571	6.631	6.756	6.479	6.582	6.608	6.540	6.518	6.552	6.827	**7.406**
	PI ↓	3.590	3.562	3.549	3.278	3.918	3.617	3.603	3.750	3.831	3.695	3.266	**3.112**
	LR-PSNR ↑	40.170	39.200	40.420	36.390	49.790	44.570	48.310	**50.900**	41.367	48.570	48.250	48.125
DIV2K	PSNR ↑	28.073	26.770	28.190	27.196	28.954	28.959	29.022	29.707	29.250	29.312	29.188	**29.857**
	SSIM ↑	0.770	0.743	0.772	0.740	0.789	0.795	0.798	0.810	0.802	0.803	0.804	**0.820**
	LPIPS ↓	0.116	0.133	0.110	0.145	0.123	0.101	0.103	0.123	**0.097**	0.103	0.104	0.111
	LPIPS-VGG ↓	0.226	0.242	0.218	0.250	0.253	**0.199**	0.212	0.237	0.202	0.200	0.210	0.213
	DISTS ↓	0.058	0.067	0.055	0.067	0.075	**0.053**	0.057	0.076	0.056	0.054	0.054	0.056
	NIQE ↓	2.953	2.911	2.952	**2.576**	3.828	2.966	3.064	3.439	3.237	3.464	2.888	2.943
	NRQM ↑	6.724	6.721	6.694	6.828	6.519	6.610	6.671	6.560	6.611	6.543	**6.870**	6.452
	PI ↓	3.137	3.126	3.158	**2.891**	3.650	3.213	3.231	3.462	3.340	3.516	3.107	3.602
	LR-PSNR ↑	42.915	38.407	42.565	37.758	50.151	45.900	50.514	**51.690**	51.088	42.950	49.057	49.076

posed wavelet guidance for the optimization objective preserves the LR manifold and generates photo-realistic high perceptual quality SR images.

Qualitative Comparison. Visual comparisons among $\times 4$ SR approaches and WGSR are presented in Fig. 2 and 5. Similar conclusions to the quantitative comparisons can be drawn from qualitative comparisons. We observe that all GAN-SR results including ESRGAN-FS [12], ESRGAN+ [50], SPSR [40], RankSRGAN [67], LDL [31] and FxSR [47] produce visible artifacts and experience excessive sharpness. On the other hand, our method WGSR is able to reconstruct the genuine image details with high reconstruction accuracy including the regions with regular patterns and the areas containing fine details such as the light pink flowers on the bush (Fig. 5). Moreover, the visual results presented in Fig. 2 certainly demonstrate the reconstruction power of WGSR when it comes to information-centric applications. The other state-of-the-art GAN-SR methods cannot recover the correct number "45". In contrast, WGSR is the clear winner for that image patch by learning to control artifacts while providing genuine image details. These improvements show that the wavelet-domain losses is a suitable optimization objective to train GAN-SR models to obtain photo-realistic, high-quality and accurate SR images.

SWT Decomposition Levels. The number of levels of the SWT decomposition is another parameter that offers flexibility on controlling genuine details vs. artifacts and affects the SR performance. The best number of levels depends on the scale and orientation of structures appearing in LR images. An example image crop containing lines with different orientation and spatial frequencies is

Figure 5. Visual comparison of the proposed wavelet-guided perceptual optimization method with the state-of-the-art methods for $\times 4$ SR on natural images from BSD100 validation set. Our method WGSR with 2-level SWT provides the best balance between perception-distortion trade-off for natural images and it has clear advantages in reconstructing realistic HF details while inhibiting artifacts. Additional visual comparisons can be found in the supplementary materials.

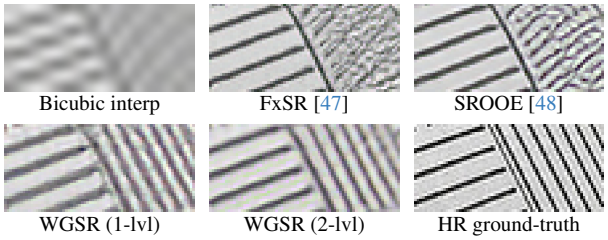

Figure 6. Visual comparison of different models on an image from Urban100 (img-92) dataset. FxSR [47], SROOE [48] and WGSR with 1-level SWT show aliasing artifacts whereas WGSR with 2-level SWT recovers all structures at different scales.

shown in Fig. 6. The state-of-the-art GAN-based SR methods FxSR [47] and SROOE [48] are unable to recover the correct structure on the right section of the crop. Here, our WGSR model with 1-level SWT can recover the correct orientiation of lines but some visible aliasing remains. The best visual result can be obtained when 2-level SWT is used by further decomposition of the LL subband of 1-level SWT into 4 subbands (L-LL, L-LH, L-HL, L-HH) and keeping the details subbands as in 1-level SWT. That is, after applying 2-level SWT, we obtain 7 subbands which brings additional flexibility in weighting loss terms for each subband. In our results, the weight parameters for 2-level decomposition is set to $\lambda_{L-LL} = 0.1$, $\lambda_{L-LH} = \lambda_{L-HL} = 0.01$ and $\lambda_{L-HH} = 0.05$, $\lambda_{LH} = \lambda_{HL} = 0.1$ and $\lambda_{HH} = 0.05$. Note that the detail (LH, HL, and HH) subbands of the 2-level decomposition are the same as in 1-level decomposition and they are assigned the same weights. We observe that computing losses on 2-level SWT manages to recover genuine details and structures when its level-2 (mid) HF subbands are penalized more in fidelity losses.

The Choice of Wavelet Family. In order to investigate the effect of the choice of wavelet family on our results, we conduct experiments with a large selection of wavelet filters including haar; db7 and db19 from Daubechies; sym7 and sym19 from Symlets; bior2.6 and bior4.4 from Biorthogonal wavelet families.

The PD trade-off performance of our WGSR model with different wavelet families on BSD100 [41] benchmark is shown in Fig. 7. We observe that the PD trade-off performance varies according to the choice of wavelet family. The best objective quality is provided by the Symlet "sym19" filter and the best perceptual quality among all solutions is achieved by the Daubechies "db7" filter. The results show that the best trade-off point is achieved by the Symlet "sym7" filter, since it is the closest to the lower left corner of the PSNR-NRQM plane. Hence, we utilize 'sym7" wavelet filter in our results.

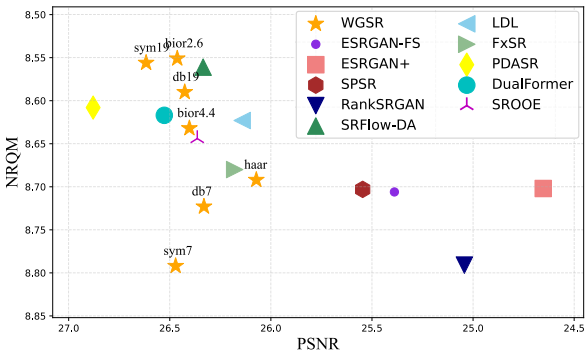

Figure 7. Perception-distortion trade-off performance of our method WGSR with different wavelet families indicated as orange stars and comparison of other state-of-the-art methods on the PSNR-NRQM plane for BSD100 [41] dataset.

4.3. Ablation Study

We conduct ablation studies to investigate the effect of each loss term in our WGSR method including the fidelity l_1, adversarial $L_{adv,G}$ and the perceptual loss L_{perc} in eqn. 2. Results are reported in Table 2. #0 refers to the baseline ESRGAN [60], where l_1 and $L_{adv,G}$ are computed in the RGB domain, and L_{perc} is taken as LPIPS.

In #1, we change L_{perc} from LPIPS [65] to DISTS [10], which results in an increase of 0.3 dB and 1.6% in objective and perceptual performance, respectively, compared to baseline #0. Similar improvements can also be observed from #4 to #5, where both l_1 and $L_{adv,G}$ are computed in the SWT domain. These results validate using DISTS instead of LPIPS navigates the SR model to a better PD point since both objective and perceptual performance improves.

Next, we investigate the effect of computing l_1 and $L_{adv,G}$ losses in the RGB vs. SWT domain. In #2, the l_1 fidelity loss is calculated in the SWT domain. We observe that the objective quality is improved by almost 1 dB without any change in perceptual quality. This clearly demonstrates the generated details can be better controlled by enforcing fidelity in the SWT subbands as opposed to RGB images. On the other hand, calculation of L_{adv} in the SWT domain in #3 favors the perceptual quality. Finally, combining all SWT domain losses in #5 (WGSR) achieves the best trade-off between objective and perceptual quality.

4.4. Discussion of Limitations

Although the proposed training method is effective in improving both fidelity and perceptual quality of SR images, there are some challenges that remain:

i) Neither the PSNR nor any quantitative perceptual scores are good indicators of visual artifacts. We demonstrate that WGSR is effective in suppressing visual artifacts in Figures 2, 5, and 6. However, this visual performance does not reflect on quantitative measures.

Table 2. Comparison of fidelity and perceptual performance according to the selection of domain of l_1 and $L_{adv,G}$ and type of L_{perc} evaluated on BSD100 [41] benchmark. The best and the second-best are marked in **bold** and underlined, respectively. We see that the proposed combination of losses (#5) achieves near 2 dB PSNR gain and better PI score compared to baseline (#0).

#	l_1	$L_{adv,G}$	L_{perc}	PSNR ↑	PI ↓
0	RGB	RGB	LPIPS	24.506	2.543
1	RGB	RGB	DISTS	24.812	2.502
2	SWT	RGB	DISTS	25.622	2.501
3	RGB	SWT	DISTS	24.746	**2.443**
4	SWT	SWT	LPIPS	<u>25.859</u>	2.466
5	SWT	SWT	DISTS	**26.331**	<u>2.453</u>

ii) Determining the best selection of weights on different SWT-domain loss terms is difficult. During our search for the best weights, we noticed decreasing weights on fidelity losses for LH and HL subbands causes fidelity scores to drop, and increasing the weight of the fidelity term on the HH subband decreases perceptual quality. On the other hand, higher λ_{adv} or λ_{perc} lead to improvement of perceptual quality at the expense of fidelity. Hence, the selection of weights leads to different PD trade-off points. In summary, while our results demonstrate training by wavelet-domain losses steers towards a better PD point, we believe there is still room for further improvements in discriminating genuine image details from artifacts.

5. Conclusion

The PD trade-off hypothesis states the impossibility of improving both fidelity and perceptual quality simultaneously beyond a theoretical limit, which is unknown in practical settings. This paper shows we can improve both fidelity (PSNR) and perceptual quality (NRQM) compared to most of the SOTA methods, while we can only improve on one with just a small compromise on the other vs. some other methods. Hence, we claim we can reach a better PD trade-off with the guidance of wavelet-domain losses. In particular, we propose a novel GAN-based SR model training method, which utilizes a weighted combinations of wavelet-domain losses. By controlling the strength of fidelity and adversarial losses according to the scale and orientation of image features in different subbands, our model is capable of learning genuine image details with high reconstruction accuracy without suffering HF artifacts and hallucinations. Extensive experiments on widely used benchmark datasets demonstrate that WGSR outperforms existing GAN-SR methods quantitatively and qualitatively, and provides better PD trade-off performance. The proposed method for adversarial training is generic in the sense that any off-the-shelf GAN-SR model can be easily plugged into this framework to benefit from wavelet guidance.

References

[1] E Agustsson and R. Timofte. NTIRE 2017 Challenge on single image super-resolution: Dataset and study. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR) Workshops*, 2017. 5

[2] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proc. of the British Machine Vision Conference*, pages 135.1–135.10, 2012. 5

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2018. 1, 2

[4] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. 2018 pirm challenge on perceptual image super-resolution. In *European Conf. Comp. Vision (ECCV) Workshops*, 2018. 5

[5] Praveen Kumar Chandaliya and Neeta Nain. Aw-gan: Face aging and rejuvenation using attention with wavelet gan. *Neural Comput. Appl.*, 35(3):2811–2825, 2022. 3

[6] Yuanqi Chen, Ge Li, Cece Jin, Shan Liu, and Thomas Li. Ssd-gan: measuring the realness in the spatial and spectral domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1105–1112, 2021. 3

[7] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE/CVF Conf. on Comp. Vision and Pattern Recog. (CVPR)*, pages 11057–11066, 2019. 1

[8] Xin Deng. Enhancing image quality via style transfer for single image super-resolution. *IEEE Signal Processing Letters*, 25(4):571–575, 2018. 3

[9] 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. *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 3076–3085, 2019. 3, 4

[10] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. on Patt Anal. and Mach. Intel.*, 44:2567–2581, 2020. 2, 5, 7, 8

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conf. Comp. Vision (ECCV)*, pages 184–199, 2014. 1

[12] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *IEEE/CVF Int. Conf. on Computer Vision Workshop (ICCVW)*, pages 3599–3608, 2019. 1, 3, 5, 6, 7

[13] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pages 2340–2349, 2021. 3

[14] Rinon Gal, Dana Cohen Hochberg, Amit Bermano, and Daniel Cohen-Or. Swagan: A style-based wavelet-driven generative model. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021. 3

[15] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog.*, pages 10021–10030, 2023. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 2

[17] Jinjin Gu, Haoming Cai, Chao Dong, et al. Ntire 2022 challenge on perceptual image quality assessment. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 951–967, 2022. 2

[18] T. Guo, H. S. Mousavi, T. H. Vu, and V. Monga. Deep wavelet prediction for image super-resolution. In *IEEE/CVF Conf. Comp. Vis. and Patt. Recog. (CVPRW)*, pages 104–113, 2017. 3

[19] H. Huang, R. He, Z. Sun, and T. Tan. Wavelet-SRnet: A wavelet-based CNN for multi-scale face super resolution. In *IEEE Int. Conf. Comp. Vis. (ICCV)*, pages 1689–1697, 2017. 3

[20] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 5197–5206, 2015. 1, 5

[21] Bjorn Jawerth and Wim Sweldens. An overview of wavelet based multiresolution analyses. *SIAM Review*, 36(3):377–412, 1994. 3

[22] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Int. Conf. Comp. Vision (ICCV)*, 2021. 3

[23] Younghyun Jo, Sejong Yang, and Seon Joo Kim. Srflow-da: Super-resolution using normalizing flow with deep convolutional block. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 5, 6, 7

[24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 2

[25] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 4

[26] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 1646–1654, 2016. 1

[27] Y. Kim and D. Son. Noise conditional flow model for learning the super-resolution space. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 424–432, 2021. 2

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. on Learning Representations, ICLR San Diego, CA, USA*, 2015. 5

[29] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 105–114, 2017. 1, 2

[30] Jingyun Liang, Jie Cao, Guolei Sun, K. Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *IEEE/CVF Int. Conf. on Comp. Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 2

[31] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 5657–5666, 2022. 1, 3, 5, 6, 7

[32] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE/CVF CVPR Workshops*, 2017. 1

[33] Fukang Liu, Mingwen Shao, Fan Wang, and Lixu Zhang. High-fidelity gan inversion by frequency domain guidance. *Computers and Graphics*, 114:286–295, 2023. 3

[34] Yunfan Liu, Qi Li, and Zhenan Sun. Attribute-aware face aging with wavelet-based generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11869–11878, 2018. 3

[35] A. Lugmayr, M. Danelljan, L. van Gool, and R. Timofte. SR-Flow: learning the super-resolution space with normalizing flow. In *European Conf. Comp. Vision (ECCV)*, pages 715–732, 2020. 2

[36] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. Ntire 2021 Learning the super-resolution space challenge. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 596–612, 2021. 2, 5

[37] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. Ntire 2022 challenge on learning the super-resolution space. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 785–796, 2022. 2, 5

[38] Xin Luo, Yunan Zhu, Shunxin Xu, and Dong Liu. On the effectiveness of spectral discriminators for perceptual quality improvement. In *ICCV*, 2023. 1, 3, 6, 7

[39] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Comput. Vis. Image Underst.*, 158:1–16, 2017. 2, 5

[40] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5, 6, 7

[41] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE Int. Conf. on Computer Vision. (ICCV)*, pages 416–423 vol.2, 2001. 5, 7, 8

[42] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Maintaining natural image statistics with the contextual loss. In *Asian Conf. Comp. Vision (ACCV)*, page 427–443, 2018. 2

[43] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3516–3525, 2021. 1

[44] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 5

[45] SeungJun Moon, Chaewon Kim, and Gyeong-Moon Park. WaGI: Wavelet-based GAN inversion for preserving high-frequency image details, 2023. 3

[46] 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 *European Conf. Comp Vis.ion (ECCV)*, page 191–207, 2020. 1

[47] Seung Ho Park, Young Su Moon, and Nam Ik Cho. Flexible style image super-resolution using conditional objective. *IEEE Access*, 10:9774–9792, 2022. 1, 3, 5, 6, 7

[48] 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 (CVPR)*, pages 1725–1735, 2023. 1, 3, 5, 6, 7

[49] Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10199–10208, 2023. 2, 3

[50] N. C. Rakotonirina and A. Rasoanaivo. Esrgan+: Further improving enhanced super-resolution generative adversarial network. In *IEEE Int. Conf. on Acoust., Speech and Signal Processing (ICASSP)*, pages 3637–3641, 2020. 2, 4, 5, 6, 7

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog.*, pages 10684–10695, 2022. 2

[52] T. Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 2019. 2

[53] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2

[54] F. Sahito, P. Zhiwen, J. Ahmed, and R. A. Memon. Wavelet-integrated deep networks for single image super-resolution. *Electronics*, 8:553, 2019. 3

[55] Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34:18126–18136, 2021. 3

[56] Lili Shen, Jie Yan, Xichun Sun, Beichen Li, and Zhaoqing Pan. Wavelet-based self-attention gan with collaborative feature fusion for image inpainting. *IEEE Trans. on Emerging Topics in Computational Intelligence*, pages 1–14, 2023. 3

[57] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conf. on Comp. Vision and Pattern Recog. (CVPR)*, pages 1874–1883, 2016. 2

[58] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *IEEE Int. Conf. on Comp. Vision (ICCV)*, pages 4809–4817, 2017. 1

[59] Jianyi Wang, Xin Deng, Mai Xu, Congyong Chen, and Yuhang Song. Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of com-

pressed video. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 3

[60] X. Wang, Ke Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *European Conf. on Comp. Vision (ECCV) Workshops*, 2018. 2, 3, 4, 5, 8

[61] Weiwen Wu, Yanyang Wang, Qiegen Liu, Ge Wang, and Jianjia Zhang. Wavelet-improved score-based generative model for medical imaging. *IEEE Transactions on Medical Imaging*, pages 1–1, 2023. 3

[62] S. Xue, W. Qiu, Fan Liu, and X. Jin. Wavelet-based residual attention network for image super-resolution. *Neurocomputing*, 382:116–126, 2020. 3

[63] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, pages 711–730, Berlin, Heidelberg, 2012. Springer. 5

[64] Jiqing Zhang, Chengjiang Long, Yuxin Wang, Haiyin Piao, Haiyang Mei, Xin Yang, and Baocai Yin. A two-stage attentive network for single image super-resolution. *IEEE Trans. on Circuits and Systems for Video Tech.*, 2021. 1

[65] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pages 586–595, 2018. 2, 5, 8

[66] Ran Zhang, Junchi Bin, Zheng Liu, and Erik Blasch. Chapter 13 - wggan: A wavelet-guided generative adversarial network for thermal image translation. In *Generative Adversarial Networks for Image-to-Image Translation*, pages 313–327. Academic Press, 2021. 3

[67] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Super resolution generative adversarial networks with learning to rank. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 44(10):7149–7166, 2021. 1, 2, 5, 6, 7

[68] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *Euro. Conf. Comp. Vision (ECCV)*, 2022. 2

[69] Y. Zhang, K. Li, Kai Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Euro. Conf. Comp. Vis. (ECCV)*, 2018. 1

[70] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 1

[71] Yuehan Zhang, Bo Ji, Jia Hao, and Angela Yao. Perception-distortion balanced admm optimization for single-image super-resolution. In *European Conf. on Comp. Vision (ECCV)*, 2022. 1, 3, 5, 6, 7

[72] Yuanbo Zhou, Wei Deng, Tong Tong, and Qinquan Gao. Guided frequency separation network for real-world super-resolution. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. Workshops (CVPRW)*, pages 1722–1731, 2020. 3