Fourier Space Losses for Efficient Perceptual Image Super-Resolution

Dario Fuoli[1] Luc Van Gool[1,2] Radu Timofte[1]
[1]Computer Vision Lab, ETH Zürich, Switzerland [2]KU Leuven, Belgium
{dario.fuoli, vangool, radu.timofte}@vision.ee.ethz.ch

Abstract

Many super-resolution (SR) models are optimized for high performance only and therefore lack efficiency due to large model complexity. As large models are often not practical in real-world applications, we investigate and propose novel loss functions, to enable SR with high perceptual quality from much more efficient models. The representative power for a given low-complexity generator network can only be fully leveraged by strong guidance towards the optimal set of parameters. We show that it is possible to improve the performance of a recently introduced efficient generator architecture solely with the application of our proposed loss functions. In particular, we use a Fourier space supervision loss for improved restoration of missing high-frequency (HF) content from the ground truth image and design a discriminator architecture working directly in the Fourier domain to better match the target HF distribution. We show that our losses' direct emphasis on the frequencies in Fourier-space significantly boosts the perceptual image quality, while at the same time retaining high restoration quality in comparison to previously proposed loss functions for this task. The performance is further improved by utilizing a combination of spatial and frequency domain losses, as both representations provide complementary information during training. On top of that, the trained generator achieves comparable results with and is $2.4\times$ *and* $48\times$ *faster than state-of-the-art perceptual SR methods RankSRGAN and SRFlow respectively.*

1. Introduction

Super-resolution (SR) deals with the problem of reconstructing the high-frequency (HF) information from a low-resolution (LR) image $x \in \mathbb{R}^{H\times W\times C}$, which are inherently lost after downsampling the high-resolution (HR) image $y \in \mathbb{R}^{rH\times rW\times C}$ due to the lower Nyquist frequency in the LR space (r denotes the scaling factor). Recent single image SR (SISR) methods [4, 17, 22, 19, 10, 14] have shown remarkable success at reconstructing the missing HF details, with emphasis on accurate restoration of the frequency content in the ground truth frames. This is typically performed with supervised training, where the ground truth images y are downsampled with a known kernel, *e.g.* bicubic, to obtain the LR input images x.

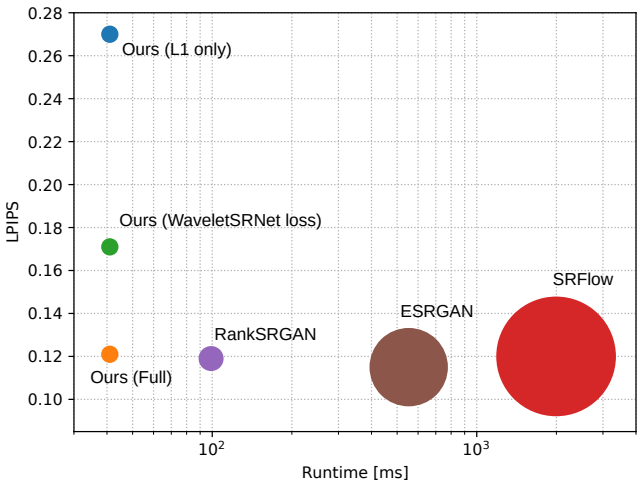

Figure 1. Runtime [ms] vs. perceptual quality (LPIPS) [36] comparison with state-of-the-art methods on DIV2K validation set. The disk area is proportional to the number of parameters. We achieve the fastest runtimes with comparable perceptual quality to much larger networks.

While it may be desirable in some applications to restore the frequencies as close to the target as possible with minimal assumptions, the ill-posed problem limits the SR networks to generate higher frequency components, as the training promotes conservative estimates imposed by the pixel-wise supervision losses. This usually results in blurry images, which appear to be of lower quality than their respective HR counterparts.

This issue has been addressed in the literature [20, 32] by employing different losses, that are designed to promote the higher frequencies for perceptually more pleasing images. These supervised objectives are often used in combination with generative adversarial networks [8] (GAN) for additional distribution learning of the HF space. Conditional GAN-based learning enables the generation of plau-

sible high frequencies without the need for strict ground truth accuracy. A lot of research has been devoted to design such perceptual losses and to find suitable combinations for pleasing results.

Recently, more and more deep learning based algorithms are implemented on smartphones, which requires low-complexity networks for fast inference and inexpensive deployment. Therefore, the design focus is slowly shifting from high-quality, high-performance methods with high-complexity networks to more efficient enhancers, which upscale faster and require less resources. In contrast to empowering a deep neural network's performance by simply increasing its complexity, which is generally straight-forward, finding an efficient network with high-performance is a much harder challenge. Searching for effective low-complexity networks with high performance, that are on par with state-of-the-art methods, is the ultimate challenge in network design.

Three main ingredients are necessary in order to maximize performance and efficiency of deep neural networks. First, the best architecture design for the task has to be determined. Usually, this task is performed manually by experts. In addition to handcrafted designs, neural architecture search algorithms [7, 6, 21] have recently been proposed to automate this task. Second, the design of the optimal loss function is imperative to fully leverage a network's performance. Third, the amount and quality of data plays a key role to maximize performance. A large portion of existing literature in SR deals with the first point. We regard the solution to the third point as straight-forward, as data can be collected efficiently for most applications. In this paper we propose a solution to the second point and try to maximize the performance of a recently proposed efficient low-complexity network [14, 35] for perceptual SR, solely by the application of our proposed loss functions.

The design of perceptual losses predominantly focuses on the spatial domain [32, 20]. However, SR is tightly coupled to the frequency domain, as only high frequencies are removed during the downsampling process. We leverage this fact and propose novel loss functions in Fourier space by calculating the frequency components with the fast Fourier transform (FFT) for direct emphasis on the frequency content. We propose a supervision loss in direct reference to the ground truth directly in Fourier domain for reconstruction. Additionally, we propose a discriminator architecture to learn the HF distribution in an adversarial training setup, working directly in Fourier space. To the best of our knowledge we are the first to apply a GAN loss directly on Fourier coefficients in SR. Our ablation study shows clear benefits over spatial losses for the task of perceptual SR. Also, employing a loss in Fourier space introduces global guidance as opposed to pixel-wise evaluation due to the nature of the Fourier transform. In order to leverage both global and local guidance, we also add the corresponding spatial supervision and GAN losses. Together with an additional perceptual loss (VGG [30]), this outperforms all other configurations in our ablation study. In addition to the advantage of our proposed losses over existing ones, we compare our trained efficient generator with high-performance state-of-the-art methods. It shows, that our losses can substantially increase the performance of a low-complexity generator to even compete with much larger networks.

2. Related Work

SR is a popular topic and a series of competitions are conducted by [31, 1, 2, 35, 33, 34, 23] which provide a broad overview of research and development over recent years in this area.

Restoration Learning based approaches have shown to be highly effective so solve the problem of SR and are therefore predominantly used in research. SRCNN [4] is one of the first convolutional neural network (CNN) based methods to surpass non-learning based SR algorithms, VDSR [17] is an improved version which adopts a deeper network for improved performance. Further concepts and improvements are explored [20, 22, 19, 10, 14] with the aim of reconstructing the missing details in a LR image as close to the ground truth as possible.

Perceptual SR Since even the best of the aforementioned methods tend to produce blurry images, another family of methods [20, 32, 37] tries to further improve the perceptual image quality by sacrificing restoration quality for increased generation of HF content [2]. For that matter, SRGAN [20] proposes the application of a generative adversarial network (GAN) [8] to better model the HF distribution in an image. The authors also propose a perceptual loss, based on features of VGG [30], which significantly boosts the perceptual quality. ESRGAN [32] extends this concept by adopting an improved GAN-loss formulation [16] and a stronger generator architecture. RankSRGAN [37] is another approach to achieve improved perceptual image quality. It uses a ranker to enable gradient based training with non-differentiable handcrafted no-reference image quality metrics. First, a dataset with pairs of images and their calculated quality score is prepared, then a ranker is trained to relatively rank two images in a differentiable manner. The learned differentiable ranker is then used in a gradient based adversarial training setup. More recently, SR-Flow [24] uses normalizing flows [28] for perceptual image SR. The method explicitly models the ambiguity in HR space and is trained by maximimum likelihood with the use of a network that is invertible by design.

Frequency-based SR Since SR is the problem of restoring frequency components, several works [12, 5, 9, 15, 3] propose to model the problem closer to frequency space in

various configurations. WaveletSRNet [12] uses wavelets to decompose the LR image by the Haar transform and generates the missing HF wavelet coefficients instead of HR images directly. Additionally, the losses are optimized for perceptual image quality by weighing the wavelet coefficients by some heuristic, in order to balance the importance of different sub-bands. DWSR [9] uses a similar approach without a weighting scheme and uses only four sub-bands, without explicit perceptual components. The loss in [12] is composed of more sub-bands, but it does not fully decompose the image as we do by applying the Fourier transform. A more recent work [15] proposes a supervision loss in Fourier space as additional loss for generative tasks. However, this work uses a different loss formulation, *i.e*. it computes the differences directly between the complex components without transformation into amplitude and phase. On top of that, to the best of our knowledge, we are the first to also employ a GAN loss directly in Fourier space.

3. Proposed Method

The task of image SR, is to increase the resolution of an image $x \in \mathbb{R}^{H \times W \times C}$ from the LR domain $\mathcal{X}$ to the corresponding image $y \in \mathbb{R}^{rH \times rW \times C}$ in HR domain $\mathcal{Y}$ with factor r. According to Nyquist–Shannon's sampling theorem, the missing HF content above the Nyquist frequency n_c must be recovered in order to get an image y from the target HR domain $\mathcal{Y}$. In contrast to the representation of an image in spatial domain, these missing frequencies can be clearly separated in Fourier domain. We therefore propose two losses in the frequency domain, to directly emphasize the training on the relevant frequencies. Additionally, the frequency components provide global guidance during training due to the nature of the Fourier transform.

3.1. Generator

Our aim is to reduce the computational complexity of the generator network for faster runtimes, while retaining the representational power for SR as high as possible. Therefore, the design of more effective losses is imperative. Improving the loss design can yield stronger gradient signals which better guide the generator during the training process. In order to test the effectiveness of our proposed losses, we use a lightweight model based on the IMDN network [14] from the same authors. This is the winner of the "AIM 2019 Challenge on Constrained SR" [35]. The network is used as an example of an efficient generator architecture to showcase the power of our loss designs against typical existing losses. The network consists of repeated information multi-distillation blocks (IMDBs), that are designed to effectively integrate information from the LR-space towards the HR-space. The whole processing is conducted in LR-space for efficiency reasons. Only in the last processing step, the refined HR image is upsampled with a standard shuffling block [29]. Generator G super-resolves a LR image $x \in \mathbb{R}^{H \times W \times C}$ into a HR image $\hat{y} = G(x) \in \mathbb{R}^{rH \times rW \times C}$.

3.2. Fourier Transform and SR

The Fourier transform is widely used to analyze the frequency content in signals. It can be applied to multi-dimensional signals such as images, where the spatial variations of pixel-intensities have a unique representation in the frequency domain. The discrete Fourier transform (DFT) decomposes an image $x \in \mathbb{R}^{H \times W \times C}$ from the spatial domain into the Fourier domain. The Fourier space is spanned by complex orthonormal basis functions, where the complex frequency components $X \in \mathbb{C}^{U \times V \times C}$ characterize the image.

$$\mathcal{F}\{x\}_{u,v} = X_{u,v} = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x_{h,w} e^{-i2\pi\left(u\frac{h}{H} + v\frac{w}{W}\right)} \tag{1}$$

Since images are composed of multiple color channels, we calculate the Fourier transform for each channel separately. The explicit notation of channels is omitted in our formulas. Each complex component $X_{u,v}$ can be represented by amplitude $|\mathcal{F}\{x\}_{u,v}|$ and phase $\angle\mathcal{F}\{x\}_{u,v}$, which provides a more intuitive analysis of the frequency content.

$$|\mathcal{F}\{x\}_{u,v}| = |X_{u,v}| = \sqrt{\mathcal{R}\{X_{u,v}\}^2 + \mathcal{I}\{X_{u,v}\}^2} \tag{2}$$

$$\angle\mathcal{F}\{x\}_{u,v} = \angle X_{u,v} = \operatorname{atan2}(\mathcal{I}\{X_{u,v}\}, \mathcal{R}\{X_{u,v}\}) \tag{3}$$

Due to symmetry in the Fourier space (Hermitian symmetry) for real valued signals x, we can omit redundant spectral components and only treat half of X, and still retain the full information in x.

$$\mathcal{F}\{x\}_{u,v} = \overline{\mathcal{F}\{x\}}_{-u,-v} \tag{4}$$

Thus, processing can be significantly reduced by neglecting redundant components when working in the Fourier domain of real-valued signals like images. Note, despite discarding the redundant values, the total number of values in the spatial and Fourier domain remains the same since the components in Fourier space are composed of real and imaginary part (or amplitude and phase).

Since the Fourier transformation assumes an infinite signal in the transformation dimensions, finite signals like images should be preprocessed to avoid edge induced artifacts. We avoid such artifacts by applying a Hann window, which suppresses the signals' amplitude towards the edges in order to smooth out the transitions. Afterwards, the image is transformed with a more accurate representation of the frequency spectrum.

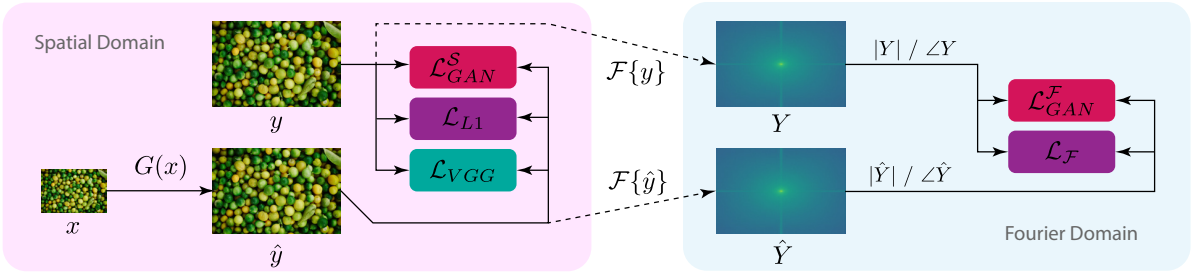

Figure 2. Overview of the proposed method. We employ losses in both spatial and Fourier domain to strengthen the training signal.

As SR is the task of reconstructing the missing HF content from a downscaled image, a reduction in the sampling rate leads to a lower Nyquist frequency n_c in the LR-space, which constitutes a hard limit in the representation capability of high frequencies above said frequency. Therefore, SR deals with the problem of generating these missing frequencies, which can be seen as the extrapolation from low to high frequencies. Contrary to the representation of an image in the spatial domain, these frequencies can be clearly separated in frequency space in order to directly emphasize the important image features for SR. Additionally, the Fourier components provide global information about the image as opposed to local information represented by pixel values in the spatial domain. We leverage these properties to design new losses for efficient perceptual SR training.

In contrast to the Fourier transform, wavelet-transforms balance spatial and frequency precision in an image by decomposing it into different sub-bands. This property is useful for many practical applications where this trade-off is inevitable. However, we are not forced to find a balance. For application in our losses, we can both leverage the frequency content with maximum precision, represented by one component for each frequency in the signal and get precise local guidance through the spatial representation of the image.

3.3. Supervision Losses

For perceptual SR, predominantly spatial domain based losses, spatial feature losses, or frequency-band separation strategies in the spatial domain, *e.g.* separation by wavelet decomposition or filtering, are proposed [32, 5]. Presumably, because most existing architectures are based on convolutions that expect spatial invariance in the input and also due to easy handling of variable image sizes with convolutional networks. Popular choices for supervision losses, *i.e.* with reference to a ground truth, are pixel-based losses L1/L2 and feature based VGG-loss [30, 20]. As proposed in [32] and for direct comparison, we investigate L1 (5) and VGG-loss (6).

$$\mathcal{L}_{L1} = \frac{1}{HW} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} |\hat{y}_{h,w} - y_{h,w}| \tag{5}$$

$$\mathcal{L}_{VGG} = \frac{1}{IJC} \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} \sum_{c=0}^{C-1} \left| N^{54}_{vgg}(\hat{y})_{i,j,c} - N^{54}_{vgg}(y)_{i,j,c} \right| \tag{6}$$

Following the setting in [32] we calculate a VGG-loss using the pre-trained 19-layer VGG network. In particular, the L1-loss between features $N^{54}_{vgg}(\cdot)$ (54 indicates 4th convolution before the 5th pooling layer) from generator output $\hat{y} = G(x)$ and the target y constitutes the VGG-loss.

In addition to these spatial domain losses, we propose a Fourier space loss $\mathcal{L}_{\mathcal{F}}$ for supervision from the ground truth frequency spectrum during training. First, ground truth y and generated image $\hat{y}$ are pre-processed with a Hann window, as described in Section 3.2. Afterwards, both images are transformed into Fourier space by applying the fast Fourier transform (FFT), where we calculate amplitude and phase of all frequency components. The L1-loss of amplitude difference $\mathcal{L}_{\mathcal{F},|\cdot|}$ and phase difference $\mathcal{L}_{\mathcal{F},\angle}$ (we take into account the periodicity) between output image and target are averaged to produce the total frequency loss $\mathcal{L}_{\mathcal{F}}$. Note, since half of all frequency components are redundant, the summation for u is performed up to $U/2-1$ only, without affecting the loss due to Eq. (4).

$$\mathcal{L}_{\mathcal{F},|\cdot|} = \frac{2}{UV} \sum_{u=0}^{U/2-1} \sum_{v=0}^{V-1} \left| |\hat{Y}|_{u,v} - |Y|_{u,v} \right| \tag{7}$$

$$\mathcal{L}_{\mathcal{F},\angle} = \frac{2}{UV} \sum_{u=0}^{U/2-1} \sum_{v=0}^{V-1} \left| \angle \hat{Y}_{u,v} - \angle Y_{u,v} \right| \tag{8}$$

$$\mathcal{L}_{\mathcal{F}} = \frac{1}{2}\mathcal{L}_{\mathcal{F},|\cdot|} + \frac{1}{2}\mathcal{L}_{\mathcal{F},\angle} \tag{9}$$

Theoretical benefits of applying a supervision loss in Fourier domain are two-fold. (1) The direct emphasis, especially on the missing HF components, promotes generation in these important areas as opposed to spatial losses (L1/L2), which are known to produce blurry images. (2) Due to the nature of the Fourier transform, which computes the frequency content with highest precision in trade-off for spatial precision, the loss provides global guidance during

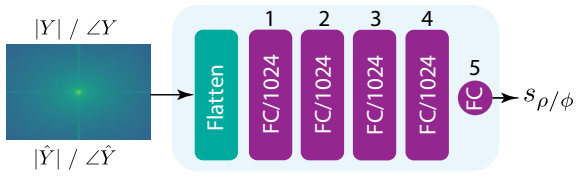

Figure 3. Proposed Fourier GAN architecture. We process the Fourier components of y and $\hat{y}$ with a fully connected network to predict real s_ρ and fake s_ϕ scores.

training in contrast to local pixel-based losses in spatial domain.

In contrast to other frequency-based losses, proposed in the literature, we directly apply the losses in Fourier space, and do not tune our losses according to some heuristic, as in [12].

3.3.1 GAN Losses

In order to further boost the perceptual quality we employ a GAN training scheme with two types of GAN-architectures, applied in spatial and Fourier domain. Learning the mapping from LR to HR directly from the ground truth severely limits the generation of images with high perceptual quality. Minimizing the risk towards a single realisation represented by the ground truth is too strict because the problem is ill-posed. A GAN training strategy relaxes the loss formulation by allowing plausible HR reconstructions resembling images from the target distribution.

We use the discriminator from [32] for our spatial GAN loss $\mathcal{L}^{\mathcal{S}}_{GAN}$. Additionally, we design a discriminator working directly in Fourier domain for our proposed frequency domain GAN-loss $\mathcal{L}^{\mathcal{F}}_{GAN}$. After the transformation of an image into Fourier space, the spatial invariance assumption is no longer valid. Therefore, the application of a convolutional architecture will not be optimal for this task. Thus, we apply a fully connected discriminator network for adversarial guidance in Fourier space, see Fig. 3. Again, generated image $\hat{y}$ and ground truth y are transformed into frequency components represented by amplitude and phase in Fourier space after the application of a Hann window.

Both adversarial losses are evaluated by a relativistic GAN formulation [16], which showed improved performance in SR over the standard GAN formulation in [32]. The discriminator's real and fake logits $s_\rho = D(y), s_\phi = D(\hat{y}) = D(G(x))$ are processed with the relativistic transformation by averaging the logits over the batch dimension b, Eq. (10), and subtracting them from the original logits, Eq. (11). The transformed real and fake scores ρ, ϕ are then evaluated with the sigmoid cross-entropy GAN-objective in (12).

$$\overline{s}_\rho = \frac{1}{B}\sum_b^B s_\rho(b), \quad \overline{s}_\phi = \frac{1}{B}\sum_b^B s_\phi(b) \tag{10}$$

$$\rho = D(y) - \overline{s}_\phi, \quad \phi = D(\hat{y}) - \overline{s}_\rho \tag{11}$$

$$\begin{aligned} \mathcal{L}^G_{GAN} &= -\frac{1}{2}\mathbb{E}_{x,y}\left[\log(\sigma(\phi)) + \log(1-\sigma(\rho))\right] \\ \mathcal{L}^D_{GAN} &= -\frac{1}{2}\mathbb{E}_{x,y}\left[\log(\sigma(\rho)) + \log(1-\sigma(\phi))\right] \end{aligned} \tag{12}$$

3.4. Training Setup

The complete training setup (13) consists of two supervision losses and two GAN-losses in both spatial and Fourier domain and an additional VGG-loss. These loss components are weighted with factors α, β, γ and minimized with Adam [18] optimizer in alternating steps.

$$\begin{aligned} &\min_G \ \alpha\left(\frac{\mathcal{L}^{G,\mathcal{S}}_{GAN} + \mathcal{L}^{G,\mathcal{F}}_{GAN}}{2}\right) + \beta\left(\frac{\mathcal{L}_{L1} + \mathcal{L}_{\mathcal{F}}}{2}\right) + \gamma\mathcal{L}_{VGG} \\ &\min_D \ \alpha\left(\frac{\mathcal{L}^{D,\mathcal{S}}_{GAN} + \mathcal{L}^{D,\mathcal{F}}_{GAN}}{2}\right) \end{aligned} \tag{13}$$

4. Experiments and Results

All settings [1] are trained on the DF2K dataset with a scaling factor of $r = 4$. DF2K is a combination of DIV2K [1] and Flickr2K [31]. Training pairs consist of paired crops of size 64×64 and 256×256 from LR and HR respectively. We evaluate all experiments on the DIV2K validation set, the standard benchmark for HR image SR. Additionally, we provide results on Urban100 [13]. For more evaluations, please refer to the supplementary material.

We calculate restoration metrics PSNR and SSIM (both on Y in YCbCr color space), perceptual metric LPIPS [36] and distributional similarity by FID [11, 27]. We deliberately refrain from using no-reference metrics, since we want to learn the image quality from the target domain, which is different to learning for a no-reference metric, as these handcrafted metrics do not necessarily correlate with the properties of the target image distribution.

4.1. Ablation

We conduct an ablation study with different loss configurations to show the effectiveness of our proposed Fourier domain losses, see Tab. 1. The generator is initialized with pretrained weights (L2) in all configurations and trained

[1] We provide codes at https://github.com/dariofuoli/FourierSpaceLosses.

Configuration	Generator	$\mathcal{L}_{L1}$	$\mathcal{L}_{F}$	$\mathcal{L}^{\mathcal{S}}_{GAN}$	$\mathcal{L}^{\mathcal{F}}_{GAN}$	$\mathcal{L}_{VGG}$	↑PSNR	↑SSIM	↓LPIPS	↓FID
1	IMDN [14]	✓					30.56	0.837	0.270	22.91
2	IMDN [14]		✓				29.53	0.811	0.189	16.98
3	IMDN [14]	✓	✓				30.32	0.834	0.266	21.96
4	IMDN [14]	✓		✓		✓	27.94	0.751	0.131	17.07
5	IMDN [14]		✓		✓	✓	29.06	0.796	0.129	17.17
6	IMDN [14]	✓	✓	✓		✓	27.96	0.762	0.127	16.94
7	IMDN [14]	✓	✓		✓	✓	29.13	0.794	0.127	17.90
8	IMDN [14]	✓	✓	✓	✓	✓	28.42	0.776	0.124	15.88
9	ESRGAN [32]	✓	✓	✓	✓	✓	28.63	0.780	0.113	14.80
10	ESRGAN [32]	✓		✓		✓	28.19	0.769	0.115	15.37

Table 1. Ablation study results. We compare different configurations of loss functions. We calculate restoration metrics PSNR and SSIM, perceptual metric LPIPS [36] and distributional similarity by FID [11]. The metrics are calculated on the DIV2K validation set.

on DF2K for 500k iterations with a constant learning rate $l = 10^{-5}$ and a batch size of $B = 16$. We do not use a learning rate scheduler for stability reasons and fairness due to the heterogeneous combinations of different loss types. The training parameters are set to $\alpha = 0.005$, $\beta = 0.01$ and $\gamma = 1$ as proposed in state-of-the-art method ESRGAN [32]. The averaging by factor 2 in (13) is removed whenever a single loss is employed per parameters α or β, to keep the balance between supervision and GAN-losses in all configurations. Additionally, we refine the pretrained generator from ESRGAN with our additional losses in the same setting with $B = 8$.

A comparison between configuration 1 and 2 clearly shows the effectiveness of our proposed Fourier domain supervision loss $\mathcal{L}_{\mathcal{F}}$ for perceptual quality enhancement. Calculating the losses with our proposed formulation significantly improves the perceptual image quality in trade-off with restoration quality [2], which is reflected by the large improvement of LPIPS (-0.081) and FID (-5.93).

Configuration 4 represents the loss formulation from ESRGAN [32], these spatial losses are exchanged by our proposed Fourier domain losses $\mathcal{L}_{\mathcal{F}}$ and $\mathcal{L}^{\mathcal{F}}_{GAN}$ in configuration 5. The perceptual quality remains comparable between the two configurations. However, the restoration quality is significantly higher compared to the ESRGAN losses by a large margin, reflected by a gain in PSNR and SSIM of +1.12dB and +0.045 respectively, which shows the superiority of our proposed Fourier domain losses over the corresponding spatial losses employed in ESRGAN.

Configuration 8 shows the effectiveness of our proposed Fourier domain losses in combination with spatial losses. It achieves the best LPIPS and FID scores of all configurations and clearly outperforms the losses of ESRGAN in configuration 4 in all metrics. Simultaneous application of losses in both spatial and frequency domain leverages complementary information from each image representation to significantly improve overall guidance during training. Configuration 9 shows the combination of ESRGAN generator with our proposed full combination of Fourier domain and spatial losses. We note the improvement (PSNR +0.44dB, FID -0.57) brought by our Fourier domain losses over the original ESRGAN in configuration 10.

4.2. Comparison with State-of-the-art

In addition to the effectiveness of our losses for perceptual performance in our ablation study, we show that we can also compete with state-of-the-art methods with a more efficient generator network, due to our better losses. We tweak the loss weights towards higher perceptual quality in trade-off with restoration quality and set them to $\alpha = 0.0025, \beta = 0.005, \gamma = 1$ for our model in Tab. 3 and Tab. 2. Note, the proposal of our losses is to showcase the improved training performance which enables to train high-performance low-complexity generators, not necessarily to achieve state-of-the-art performance. Despite the low complexity of G in our setting, we are able to compete with image quality of state-of-the-art methods, with a substantial reduction of runtime.

ESRGAN uses a combination of L1, VGG and GAN loss and proposes an improved generator architecture derived from SRGAN [20]. **RankSRGAN** [37] introduces a method to use non-differentiable handcrafted image quality metrics (Ma [25], NIQE [26] and PI [2]) for training in a GAN-based setup. The generator network in RankSRGAN is SRGAN [20]. **SRFlow** [24] is a recently proposed method, which uses normalizing flows [28] for perceptual image SR. The concept of normalizing flows provides an alternative to GAN-based learning by modeling the ill-posed problem explicitly as a stochastic process. We also compare our loss formulation to recently proposed losses using the wavelet transformation, see Sec. 3.2. Division into subbands with wavelet transform is used by **WaveletSRNet** [12] and DWSR [9], which both use the Haar transform. We compare our method to the losses in Wavelet-

SRNet which uses a finer division and a more sophisticated loss formulation than DWSR. For this purpose, we train the efficient generator backbone G with the proposed losses in WaveletSRNet for direct comparison.

For all other methods we use the pretrained models provided by the authors, as all of them are trained on DF2K. Additionally, we provide the results for standard bicubic upsampling as a baseline. To quantify model complexity and efficiency, we compute number of parameters and runtimes at inference on a NVIDIA TITAN RTX and an Intel i7 CPU (6 cores). We also provide visual examples in Fig. 4 which support our quantitative evaluation.

4.2.1 Discussion

The superiority of our losses compared to ESRGAN's losses is already shown in the ablation study in Tab. 1. On top of that, we can even compete with ESRGAN's high-complexity generator, which achieves slightly better LPIPS and FID values, but lower PSNR and SSIM scores with a substantially slower inference time by a factor of over $13\times$ on GPU.

Our losses significantly surpass all three RankSRGAN models in both restoration metrics PSNR/SSIM and even achieve the highest FID score. Only the NIQE and PI optimized models have slightly higher LPIPS scores, which however comes with a $2.4\times$ higher runtime on GPU. Note, this is a substantial difference in complexity, *e.g.* this equates to reducing the number of layers in a network by a factor of 2.4. In comparison to the ranker approach, our loss formulation does not depend on the difficult design of a meaningful handcrafted quality metric, which manifests an upper bound on the achievable quality. We also do not require the expensive setup of the ranker, we achieve stronger guidance by direct emphasis on the frequency content without an additional explicit concept of perceptual quality.

SRFlow [24] is the most expensive method with a large number of parameters and slow inference speeds of 1.995s and 55.33s on GPU and CPU respectively, yet does not outperform the performance of other methods, with the exception of PSNR and SSIM. Our highly efficient method is on par with SRFlow with comparable perceptual metrics. Our solution has better FID score (+0.41) but slightly lower LPIPS score (-0.001). However, there is an enormous difference in inference speed, *e.g.* SRFlow is 48 times slower on GPU than our method, which highlights the superiority of our proposed losses.

We train G with WaveletSRNet's losses from scratch with a learning rate of $l = 10^{-5}$ for 500k iterations with a batch size of $B = 16$. Further, we finetune G with a lower learning rate of $l = 10^{-6}$ for another 250k iterations. We substantially outperform WaveletSRNet's loss formulation with our proposed losses in regard to PSNR and perceptual metrics LPIPS and FID. Our proposed supervision loss in Fourier space $\mathcal{L}_{\mathcal{F}}$, from our ablation study, already outperforms WaveletSRNet in three metrics with the exception of LPIPS, see configuration 2 in Tab. 1.

Method	↑PSNR	↑SSIM	↓LPIPS	↓FID
ESRGAN (Our losses) [32]	25.05	0.738	0.120	24.07
ESRGAN [32]	24.36	0.717	0.123	25.50
RankSRGAN (NIQE) [37]	24.52	0.715	0.143	27.47
Ours (Full)	24.69	0.723	0.132	26.70

Table 2. Evaluation on Urban100. Red indicates best, blue second best.

We evaluate our losses on Urban100 [13] in Tab. 2 to show the generalization capability of our approach. Our efficient setting achieves comparable performance also on this dataset. The application of our losses to ESRGAN again results in clear improvements in all 4 metrics by a substantial margin, especially in the restoration metrics.

5. Conclusion

We present two Fourier domain losses – a supervision and a GAN loss – to strengthen the training signal for the task of perceptual image SR. Our ablation study shows the provision of complementary information during training in addition to the losses in spatial-domain. Due to the improved guidance, it is possible to train a significantly lower complexity – and therefore faster – network to achieve comparable performance of much larger networks, which we regard as an important property for many practical applications. The runtime of the generator backbone can be cut down to only 41ms, which is over $13\times$ faster than ESRGAN and $48\times$ faster than SRFlow on GPU. The separation of images into LF and HF content and therefore the direct emphasis on the missing high frequencies in Fourier space, imposed by our losses, helps the SR network to generate plausible HF content. At the same time, we also apply the corresponding spatial losses to leverage the complementary local information, which results in even better perceptual quality. To the best of our knowledge, we are the first to successfully apply a GAN-based loss directly on Fourier components for SR. We are convinced that more research into architectural improvements of our Fourier-space GAN-network can further advance the effectiveness of our approach.

Acknowledgements.
This work was partly supported by a Huawei Technologies Co. Ltd project and the ETH Zürich Fund (OK).

Method	↑PSNR	↑SSIM	↓LPIPS	↓FID	↓Par [M]	↓GPU [s]	↓CPU [s]
Bicubic	28.11	0.782	0.410	44.79	-	-	-
SRFlow [24]	28.68	0.773	0.120	16.13	39.542	1.995	55.33
ESRGAN [32]	28.19	0.769	0.115	15.37	16.698	0.553	29.28
RankSRGAN (Ma) [37]	27.30	0.742	0.141	18.40	1.554	0.099	3.97
RankSRGAN (NIQE) [37]	28.19	0.765	0.119	15.89	1.554	0.099	3.97
RankSRGAN (PI) [37]	28.11	0.765	0.121	16.28	1.554	0.099	3.97
Ours (WaveletSRNet loss [12])	27.97	0.786	0.171	19.80	0.894	0.041	1.72
Ours (L1 only)	30.56	0.837	0.270	22.91	0.894	0.041	1.72
Ours (Full)	28.28	0.770	0.121	15.72	0.894	0.041	1.72

Table 3. Comparison with state-of-the-art methods. We compare in terms of image quality scores (PSNR, SSIM, LPIPS and FID) and efficiency measures (parameters and runtimes). Red indicates best, blue second best.

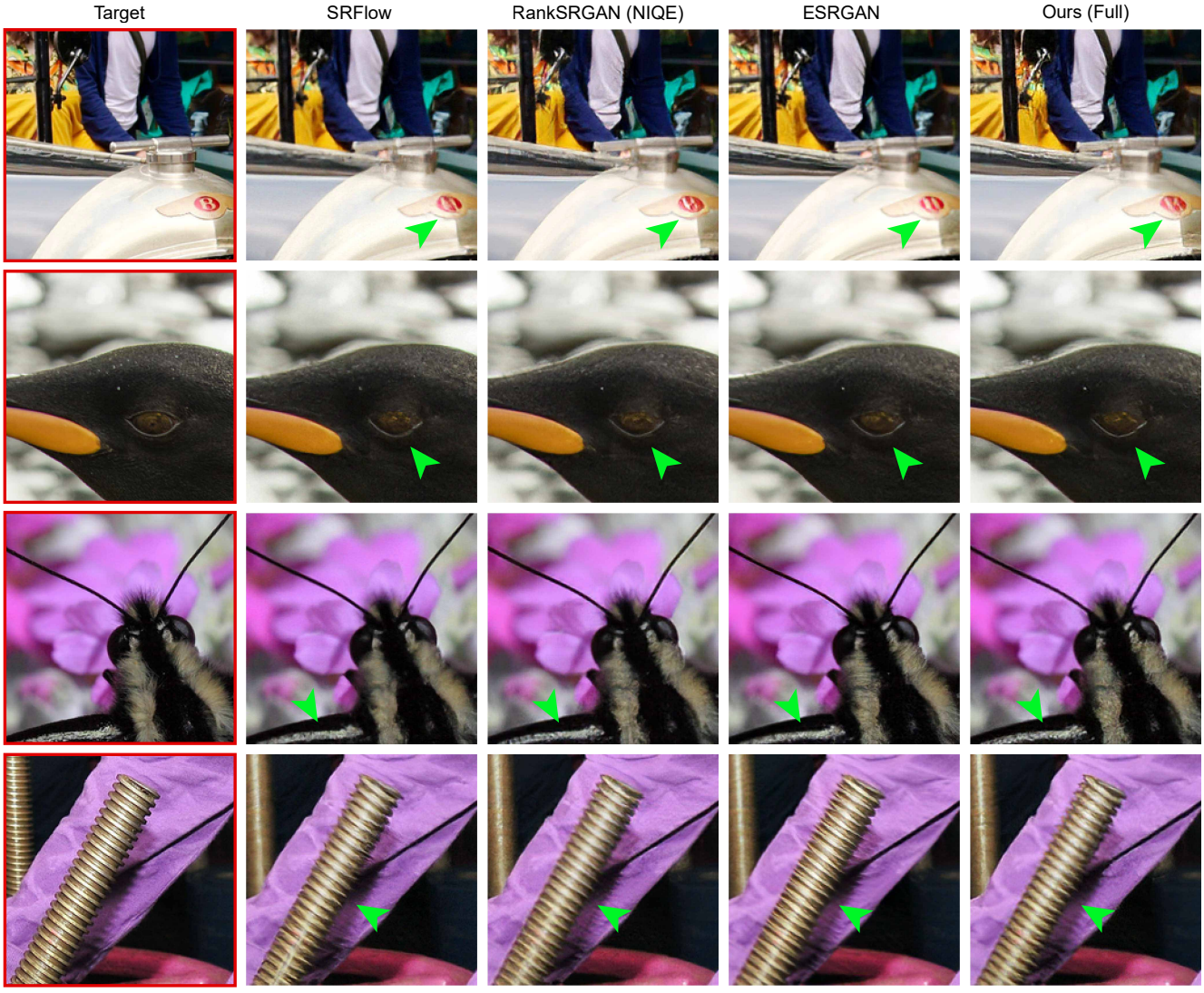

Figure 4. Visual examples from DIV2K validation images.

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 (CVPR) Workshops*, July 2017. 2, 5

[2] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018. 2, 6

[3] 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 (ICCV)*, October 2019. 2

[4] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, Feb 2016. 1, 2

[5] M. Fritsche, S. Gu, and R. Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3599–3608, 2019. 2, 4

[6] Yonggan Fu, Wuyang Chen, Haotao Wang, Haoran Li, Yingyan Lin, and Zhangyang Wang. AutoGAN-distiller: Searching to compress generative adversarial networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3292–3303. PMLR, 13–18 Jul 2020. 2

[7] Xinyu Gong, Shiyu Chang, Yifan Jiang, and Zhangyang Wang. Autogan: Neural architecture search for generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[8] 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 27*, pages 2672–2680. 2014. 1, 2

[9] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 2, 3, 6

[10] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2

[11] 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. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637. Curran Associates, Inc., 2017. 5, 6

[12] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *IEEE International Conference on Computer Vision*, pages 1689–1697, 2017. 2, 3, 5, 6, 8

[13] 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. 5, 7

[14] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia (ACM MM)*, pages 2024–2032, 2019. 1, 2, 3, 6

[15] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for generative models. *arXiv preprint arXiv:2012.12821*, 2020. 2, 3

[16] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 2, 5

[17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[19] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2

[20] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 4, 6

[21] Royson Lee, Łukasz Dudziak, Mohamed Abdelfattah, Stylianos I. Venieris, Hyeji Kim, Hongkai Wen, and Nicholas D. Lane. Journey towards tiny perceptual super-resolution. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[22] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 1, 2

[23] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[24] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *ECCV*, 2020. 2, 6, 7, 8

[25] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for

single-image super-rolution. *Computer Vision and Image Understanding*, pages 1–16, 2017. 6

[26] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 6

[27] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. Version: 0.2.0, DOI: 10.5281/zenodo.3786540. 5

[28] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. 2, 6

[29] 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 *CVPR*, 2016. 3

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 2, 4

[31] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 2, 5

[32] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018. 1, 2, 4, 5, 6, 7, 8

[33] Kai Zhang, Martin Danelljan, Yawei Li, and Radu Timofte. Aim 2020 challenge on efficient super-resolution: Methods and results. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 5–40, Cham, 2020. Springer International Publishing. 2

[34] Kai Zhang, Shuhang Gu, and Radu Timofte. Ntire 2020 challenge on perceptual extreme super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[35] Kai Zhang, Shuhang Gu, Radu Timofte, et al. Aim 2019 challenge on constrained super-resolution: Methods and results. In *ICCV Workshops*, 2019. 2, 3

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1, 5, 6

[37] 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 (ICCV)*, October 2019. 2, 6, 7, 8