# Wavelet Domain Style Transfer for an Effective Perception-distortion Tradeoff in Single Image Super-Resolution

Xin Deng
Imperial College London
x.deng16@imperial.ac.uk

Ren Yang
ETH Zurich
reyang@ee.ethz.ch

Mai Xu
Beihang University
maixu@buaa.edu.cn

Pier Luigi Dragotti
Imperial College London
p.dragotti@imperial.ac.uk

## Abstract

*In single image super-resolution (SISR), given a low-resolution (LR) image, one wishes to find a high-resolution (HR) version of it which is both accurate and photo-realistic. Recently, it has been shown that there exists a fundamental tradeoff between low distortion and high perceptual quality [3], and the generative adversarial network (GAN) is demonstrated to approach the perception-distortion (PD) bound effectively. In this paper, we propose a novel method based on wavelet domain style transfer (WDST), which achieves a better PD tradeoff than the GAN based methods. Specifically, we propose to use 2D stationary wavelet transform (SWT) to decompose one image into low-frequency and high-frequency sub-bands. For the low-frequency sub-band, we improve its objective quality through an enhancement network. For the high-frequency sub-band, we propose to use WDST to effectively improve its perceptual quality. By feat of the perfect reconstruction property of wavelets, these sub-bands can be re-combined to obtain an image which has simultaneously high objective and perceptual quality. The numerical results on various datasets show that our method achieves the best trade-off between the distortion and perceptual quality among the existing state-of-the-art SISR methods.*

## 1. Introduction

Single image super-resolution (SISR) aims to restore a high-resolution (HR) image from a low-resolution (LR) one. In this context, some methods focus on improving the objective image quality, through minimizing the mean squared error (MSE) between the restored and the ground-truth images [6, 22, 12, 13, 15, 30, 31]. Other methods aim to improve the perceptual image quality, through minimizing the perceptual loss using adversarial training [14, 21, 19]. The methods driven by objective quality can achieve low distortion but with poor perceptual quality, while the other category can generate photo-realistic images but with large MSE distortion. We wish to obtain a super-resolved image which is both accurate and photo-realistic. However, as pointed out in [3], there exists a tradeoff between the ability to achieve low MSE and high perceptual quality.

A natural approach to achieve this tradeoff is to train a generative adversarial network (GAN) to minimize a combined MSE and adversarial loss, which has been tried by both SRGAN-MSE [14] and ENet [21]. However, the training process is extremely unstable. On the one hand, the adversarial loss encourages the synthesis of high-frequency details in the results [21]. On the other hand, since these high-frequency details are not in the right place, the MSE distortion is increased. This unstable training may lead to many undesirable artifacts in the restored image, as shown in Fig. 1. To avoid this, ESRGAN [27], which is the winner of the PIRM challenge [2], proposed to train two separate networks with the low MSE and high perceptual quality targets, respectively. The two networks are then interpolated to achieve a compromise on the objective and perceptual quality. However, the network interpolation requires that the two networks have exactly the same architectures, which strongly limits their performance. Instead of the network interpolation, the image fusion method can be more flexible, since it has no constraint on the network structure. Given one image with high objective quality and another image with high perceptual quality, image fusion aims to fuse them to obtain an image with both high objective and perceptual quality. Recently, Deng [5] proposed to combine the two images using image style transfer. However, since the style transfer is performed in pixel domain, it is difficult to preserve the structure and texture information. As shown in Fig. 1, the structure of the wall is severely affected.

Another disadvantage of Deng [5] is that it tries to op-

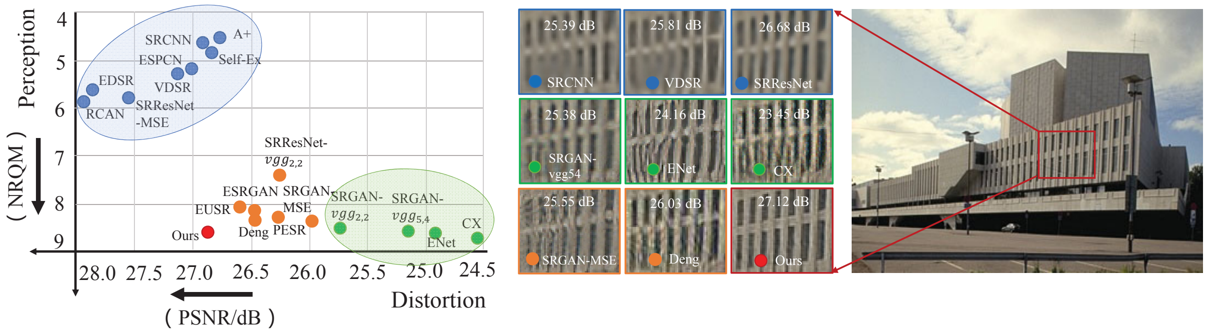



Figure 1. Perception-distortion performance of different SISR methods. The blue points represent methods aiming for objective quality, the green points represent methods aiming for perceptual quality, and the orange points represent methods aiming for a trade-off between perception and distortion. The higher PSNR value indicates better objective quality and the higher perceptual score indicates better perceptual quality. The bottom left corner is the best. Our method achieves the best trade-off among all the "orange" methods.

timize the objective and perceptual quality as a whole, but the objective and perceptual quality are influenced by different elements in an image. When they are optimized as a whole, the increase of objective quality may lead to the decrease of perceptual quality, and vice versa. To achieve the best tradeoff, we should separate the elements affecting the objective quality from those affecting the perceptual quality, and optimize each of them separately. In this paper, we propose to use wavelet transform to achieve this separation, since wavelet can split an image into one low-frequency and several high-frequency sub-bands. We find that the low-frequency sub-band plays an important role in the objective quality, while the high-frequency sub-bands can affect the perceptual quality significantly. After separation, to obtain the best tradeoff, we use an enhancement network to improve the objective quality of the low-frequency sub-band, and wavelet domain style transfer to improve the perceptual quality of the high-frequency sub-bands.

Note that in this paper, we are not aiming for a new SISR method towards high perceptual or objective image quality, which has been extensively explored recently. Instead, we propose a novel image fusion method which combines two images to achieve the best tradeoff between the perception and distortion, as shown in Fig. 1. Our method overcomes many drawbacks of the existing methods. For example, compared with SRGAN-MSE [14], we do not need to train a deep network, and thus we have no concerns on the stability of training. Compared with ESRGAN [27], we are more flexible with the choice of the network architecture, which gives us more freedom to achieve the best PD tradeoff. Compared with Deng [5], we split the elements affecting the objective quality from those affecting the perceptual quality, and we perform the style transfer in the wavelet domain with new techniques. All these contribute to higher reconstruction performance and a better PD tradeoff.

The main contributions of this work are as follows:

- We show the relationship between the objective/perceptual image quality and the wavelet subbands, which lays an important foundation to push forward the PD performance. Through the wavelet separation, the objective and perceptual quality is allowed to be enhanced separately, with little influence on the other, which leads to a better PD tradeoff.

- We propose a wavelet domain style transfer (WDST) algorithm with a new defined loss function, to achieve an effective tradeoff between distortion and perception. To the best of our knowledge, we are the first to apply style transfer in the wavelet domain towards a good PD tradeoff in SISR.

- We test the performance of our method on various datasets. Compared with other state-of-the-art methods, our method achieves a better tradeoff between the objective and perceptual quality.

## 2. Related work

**SISR methods for objective quality.** To improve the objective quality, most methods try to minimize the MSE loss between the reconstructed image and the ground-truth. Traditional methods rely on dictionary learning to learn the mapping from LR patches to HR patches [28, 29, 25]. The state-of-the-art methods trained a specially-designed deep neural network to minimize the MSE loss between the LR and HR images[6, 12, 22, 15, 8, 31, 30]. This kind of methods can generate HR images with high objective quality. However, these images are often visually unpleasant with blurred edges, due to the absence of high-frequency details, especially for large upscaling factors.

**SISR methods for perceptual quality.** Since the MSE loss cannot measure the perceptual similarity between two images, Ledig *et.al* [14] proposed to minimize the perceptual loss which was defined as a weighted sum of VGG loss and adversarial loss. The VGG loss is good at representing the perceptual similarity between two images, and the adversarial loss can make the restored image look realistic. Later, Saggadi *et.al* [21] proposed to add a texture matching loss to the VGG loss and adversarial loss, which achieved good results in reconstructing images with high perceptual quality. Recently, Mechrez *et.al* [19] proposed the contextual loss to make the internal statistics of the restored image similar to the ground-truth, which leads to more realistic images.

**SISR methods for tradeoff between objective and perceptual quality.** Both [14] and [21] have tried to optimize the objective and perceptual quality simultaneously. Specifically, in [14], the SRGAN-MSE method is proposed to minimize the combined loss of MSE and adversarial losses. In [21], another texture matching loss is added to the MSE and adversarial loss to make the training process more stable. However, their results still suffer from blocking and noisy artifacts. Choi *et.al* [4] trains a multi-scale super-resolution model with a discriminator network and two qualitative score predictors, which achieves high perceptual quality while preserving the objective quality. Most recently, ESRGAN [27] proposed to train two networks which aim to enhance the objective and perceptual quality, respectively, and then these two networks are interpolated to achieve a tradeoff between the objective and perceptual quality. The work most related with ours is [5], which also uses style transfer to combine two images. However, in [5], the style transfer algorithm is performed in the pixel domain, and it has no technique to split the objective and perceptual quality related elements from each other. As a result, the objective and perceptual quality are optimized as a whole, which significantly decreases the perception-distortion performance.

## 3. Proposed method

**Stationary wavelet transform.** The wavelet transform allows the multi-resolution analysis of images [10]. The classical discrete wavelet transform (DWT) has a drawback, i.e., it is not shift-invariant. The stationary wavelet transform (SWT), also known as undecimated wavelet transform, overcomes this drawback by removing the downsampling operation in DWT [24]. Fig. 2 illustrates the 2D SWT process for 2 level decomposition. Suppose that $H_0$ and $G_0$ are the low-pass and high-pass filters of a standard 1D wavelet decomposition, we can obtain the $z$ transform of $LL$, $LH$, $HL$, and $HH$ sub-bands at the $i$-th level through the following formulations:

$$LL_i(z_x, z_y) = H_0(z_y^{2^{i-1}})H_0(z_x^{2^{i-1}})LL_{i-1}(z_x, z_y), \quad (1)$$

$$LH_i(z_x, z_y) = G_0(z_y^{2^{i-1}})H_0(z_x^{2^{i-1}})LL_{i-1}(z_x, z_y), \quad (2)$$

$$HL_i(z_x, z_y) = H_0(z_y^{2^{i-1}})G_0(z_x^{2^{i-1}})LL_{i-1}(z_x, z_y), \quad (3)$$

$$HH_i(z_x, z_y) = G_0(z_y^{2^{i-1}})G_0(z_x^{2^{i-1}})LL_{i-1}(z_x, z_y), \quad (4)$$

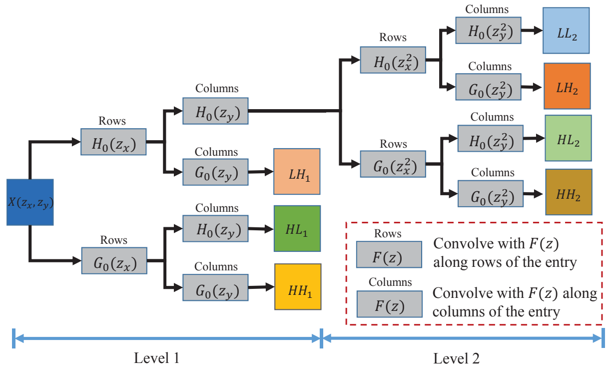


Figure 2. Illustration of two level 2D stationary wavelet transform (SWT) of image $X$, with $H_0$ and $G_0$ as the low-pass and high-pass filters, respectively.

where the $LL_{i-1}$ is the LL sub-band at the $(i-1)$-th level, with $LL_0$ as the input image $X$. After the $N$-th level decomposition, we obtain $(3N+1)$ wavelet sub-bands with the same size as the input image, i.e., $LL_N, \{LH_i\}_{i=1}^N, \{HL_i\}_{i=1}^N, \{HH_i\}_{i=1}^N$, where $LL_N$ contains the low-frequency information at the $N$-th level, $LH_i$, $HL_i$ and $HH_i$ contain the horizontal, vertical and diagonal details at the $i$-th level, respectively.

**Motivation.** The 2D SWT can decompose an image into multiple sub-bands, including one low-frequency and several high-frequency sub-bands. Our key insight here is that the low-frequency sub-band has a significant effect on the objective quality of the image, while the high-frequency sub-bands affect the perceptual quality significantly. To verify that, we consider two super-resolved images: $A_p$ with high perceptual quality but low objective quality, and $A_o$ with high objective quality but low perceptual quality. Fig. 3 shows these two images, together with the histograms of their sub-bands after SWT. Here, $A_p$ and $A_o$ are obtained using the existing SISR methods CX [19] and EDSR [15], respectively. We use peak signal-to-noise ratio (PSNR) to measure the objective quality, and NRQM [16] to measure the perceptual quality following [19]. Note that larger PSNR and NRQM values indicate better objective and perceptual quality, respectively. As shown in Fig. 3, the high-frequency sub-bands (i.e., LH, HL, HH) of $A_p$ have quite similar histogram distributions as the ground-truth, but that is not the case for $A_o$. Since the high-frequency sub-bands contain the detail information, this can explain why $A_p$ has high perceptual quality. For the LL sub-band, $A_o$ has a more similar histogram as the ground-truth than $A_p$, which is one of the reasons why $A_o$ has high objective quality.

In order to further verify our observation, a simple substitution experiment is performed as follows. We replace the low-frequency sub-band of $A_p$ with that of $A_o$, and keep all its high-frequency sub-bands. These sub-bands are combined via 2D inverse SWT (ISWT) to obtain a reconstructed image $\tilde{A}_p$. Likewise, we replace the low-frequency sub-

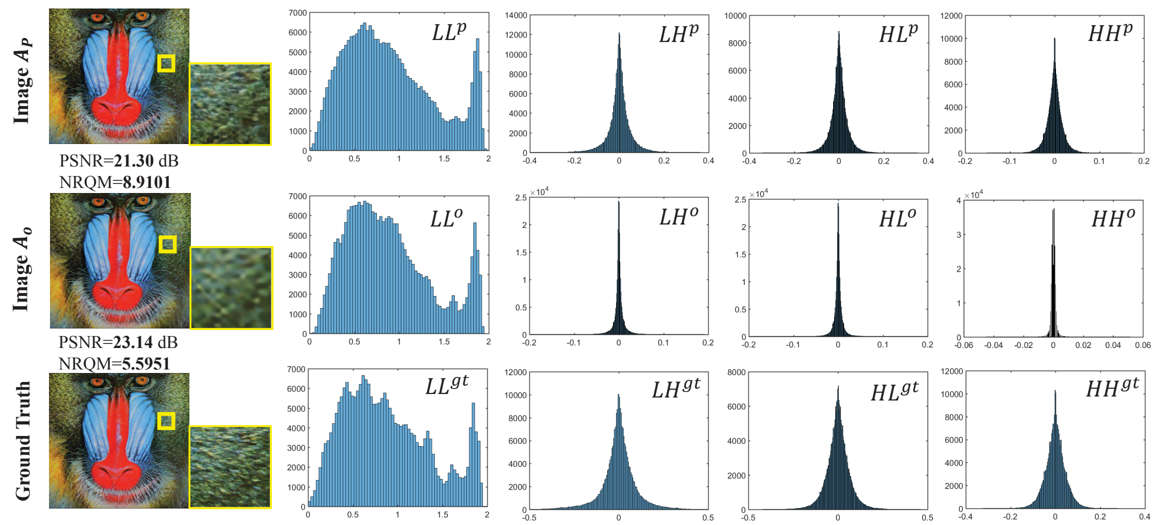
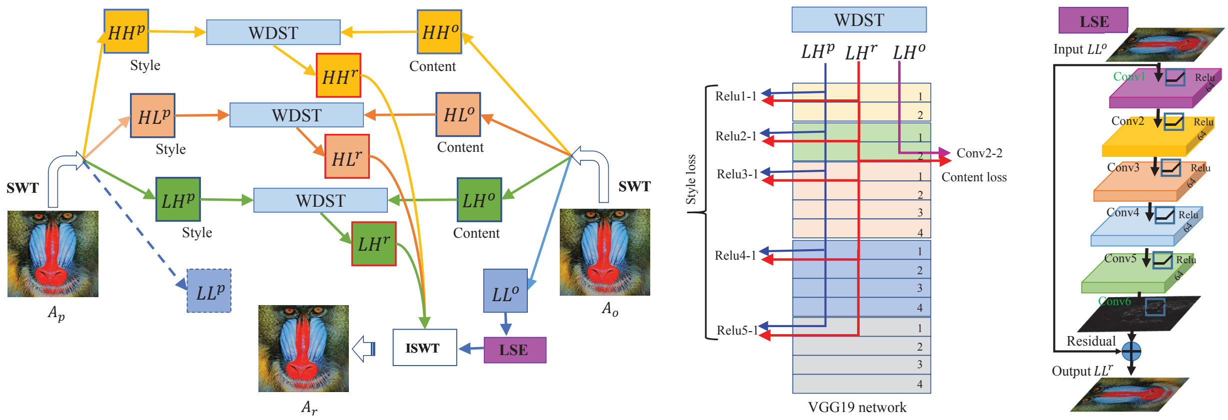



Figure 3. The first row shows the histograms of different sub-bands of $A_p$ which has high perceptual quality but low objective quality. The second row shows the histograms of different sub-bands of $A_o$ which has high objective quality but low perceptual quality. The third row shows the ground-truth histograms.




(a) (b) (c)

Figure 4. (a) shows the framework of our method, (b) illustrates the wavelet domain style transfer (WDST) algorithm, and (c) shows the low-frequency sub-band enhancement (LSE) network.

band of $A_o$ with that of $A_p$ to obtain a reconstructed image $\tilde{A}_o$. Table 1 shows the PSNR and NRQM results on the BSD100 dataset. As can be seen, the PSNR of $\tilde{A}_p$ improves more than 1dB over $A_p$ while the NRQM score does not change too much. Similar phenomenon can be observed between $\tilde{A}_o$ and $A_o$. The reason why the objective quality is significantly affected is that the low-frequency sub-band is changed. In contrast, the perceptual quality is not particularly influenced because we preserve the high-frequency sub-bands. Thus, in order to obtain an image with a good PD tradeoff, one possible solution is to pursue high objective quality of its low-frequency sub-band and high perceptual quality of its high-frequency sub-bands.

Fig. 4 (a) shows the framework of our method. Given one image $A_p$ with high perceptual quality and another image $A_o$ with high objective quality, we first perform 2D SWT on these two images, so that each image is decomposed into one low-frequency and several high-frequency sub-bands. Take the decomposition with one level for example, $A_p$ is decomposed into $\{LL^p, LH^p, HL^p, HH^p\}$, and $A_o$ is decomposed into $\{LL^o, LH^o, HL^o, HH^o\}$. For $LL_o$, we use LSE network to enhance its objective quality. For high-frequency sub-bands pairs, e.g., $LH_p$ and $LH_o$, we use WDST to fuse them to a new sub-band. Finally,

Table 1. PSNR and NRQM scores on the BSD100 dataset.

| Methods | $A_p$ | $\tilde{A}_p$ | $A_o$ | $\tilde{A}_o$ |
|---|---|---|---|---|
| PSNR | 24.58 | 25.68 | 27.80 | 26.57 |
| NRQM | 8.8007 | 8.7775 | 5.7159 | 5.8864 |

Table 2. Benchmark comparisons for 4 × upscaling, with the best results bold and the second bests underlined.

| Set5 | Bicubic | EDSR[15] | CX[19] | SRGAN-MSE[14] | G–MGBP[20] | PESR[26] | Deng[5] | ESRGAN[27] | Ours |
|---|---|---|---|---|---|---|---|---|---|
| PSNR | 28.42 | 32.63 | 29.10 | 30.66 | 30.87 | 30.76 | 31.14 | 31.11 | **31.46** |
| SSIM | 0.8245 | 0.9117 | 0.8523 | 0.8758 | 0.8807 | 0.8915 | 0.8917 | 0.8839 | **0.8929** |
| NRQM | 3.7624 | 5.2106 | 7.9566 | 7.3082 | 7.3115 | 7.1344 | 7.0022 | 7.0724 | **7.5180** |
| Set14 | Bicubic | EDSR[15] | CX[19] | SRGAN-MSE[14] | G–MGBP[20] | PESR[26] | Deng[5] | ESRGAN[27] | Ours |
| PSNR | 26.10 | 28.95 | 26.01 | 27.01 | 27.56 | 27.57 | 27.77 | 27.53 | **28.07** |
| SSIM | 0.7850 | 0.8583 | 0.7839 | 0.8033 | 0.8206 | 0.8322 | 0.8325 | 0.8228 | **0.8356** |
| NRQM | 3.6598 | 5.3788 | 7.9423 | **7.8770** | 7.5042 | 7.5301 | 7.5575 | 7.5936 | 7.6827 |
| BSD100 | Bicubic | EDSR[15] | CX[19] | SRGAN-MSE[14] | G–MGBP[20] | PESR[26] | Deng[5] | ESRGAN[27] | Ours |
| PSNR | 25.96 | 27.80 | 24.58 | 25.98 | 26.59 | 26.33 | 26.46 | 26.44 | **26.82** |
| SSIM | 0.6675 | 0.7432 | 0.6432 | 0.6429 | 0.6926 | 0.6980 | 0.7048 | 0.7002 | **0.7058** |
| NRQM | 3.7207 | 5.7159 | 8.8007 | 8.4276 | 8.1790 | 8.3298 | 8.4452 | 8.3034 | **8.5948** |
| Urban100 | Bicubic | EDSR[15] | CX[19] | SRGAN-MSE[14] | G–MGBP[20] | PESR[26] | Deng[5] | ESRGAN[27] | Ours |
| PSNR | 23.14 | 26.86 | 24.00 | - | 25.15 | 25.88 | 25.96 | 26.08 | **26.26** |
| SSIM | 0.9011 | 0.9679 | 0.9313 | - | 0.9495 | 0.9610 | 0.9620 | 0.9624 | **0.9649** |
| NRQM | 3.4412 | 5.3365 | 6.7982 | - | 6.2190 | 6.3190 | 6.4317 | 6.1762 | **6.4556** |
| PIRM | Bicubic | EDSR[15] | CX[19] | SRGAN-MSE[14] | G–MGBP[20] | PESR[26] | Deng[5] | ESRGAN[27] | Ours |
| PSNR | 26.51 | 28.72 | 25.41 | - | 27.17 | 27.11 | 27.48 | 26.66 | **27.63** |
| SSIM | 0.8232 | 0.8930 | 0.8177 | - | 0.8524 | 0.8649 | 0.8728 | 0.8529 | **0.8755** |
| NRQM | 3.8376 | 5.7116 | 8.5746 | - | 8.0556 | 8.2172 | 8.1665 | 8.2445 | **8.3692** |

all fused sub-bands and enhanced $LL_o$ are synthesised by ISWT to obtain image $A_r$.

**Low-frequency sub-band enhancement (LSE).** For the low-frequency sub-band $LL^o$, we aim to further improve its objective quality. Here, we employ the basic network structure of VDSR [12] to achieve this goal, as shown in Fig. 4 (c). The network is composed of 6 convolutional layers with a rectified linear unit (Relu) after each layer. For each layer, the filter size is $3 \times 3$ and the number of filters is 64. The input to the network is the low-frequency sub-band $LL^o$ from the image $A_o$, and the target is the $LL^{gt}$ from the ground-truth image $A_{gt}$. To speed up the training process, we also use the residual learning strategy which learns the difference between target $LL^{gt}$ and the input $LL^o$. The training goal is to minimize the $\ell_2$ norm between the predicted outputs $LL^r$ and the ground truth $LL^{gt}$:

$$\mathcal{L} = \sum_{i=1}^{N} \|LL^{gt}(i) - LL^r(i)\|_2, \tag{5}$$

where $LL^r$ is the sum of $LL^o$ and the learned residual map.

**Wavelet domain style transfer (WDST).** For the high-frequency sub-bands, we propose a wavelet domain style transfer (WDST) algorithm to improve their perceptual quality. Take the sub-band pair $LH^p$ and $LH^o$ for example, as shown in Fig. 3, the wavelet coefficients in $LH^p$ are richer than those in $LH^o$, i.e., $LH^p$ contains more nonzero wavelet coefficients than $LH^o$. We wish to transfer the detailed wavelet coefficients in $LH^p$ to $LH^o$, so that $LH^o$ can have higher perceptual quality. Thus, we regard $LH^p$ as the style input and $LH^o$ as the content input to generate an output sub-band $LH^r$ using WDST. Different from the conventional style transfer algorithm where the inputs are pixel values, we use the wavelet coefficients as inputs in the WDST. Since the wavelet coefficients can be negative or larger than 1, a pre-processing step is required to normalize them between 0 and 1.

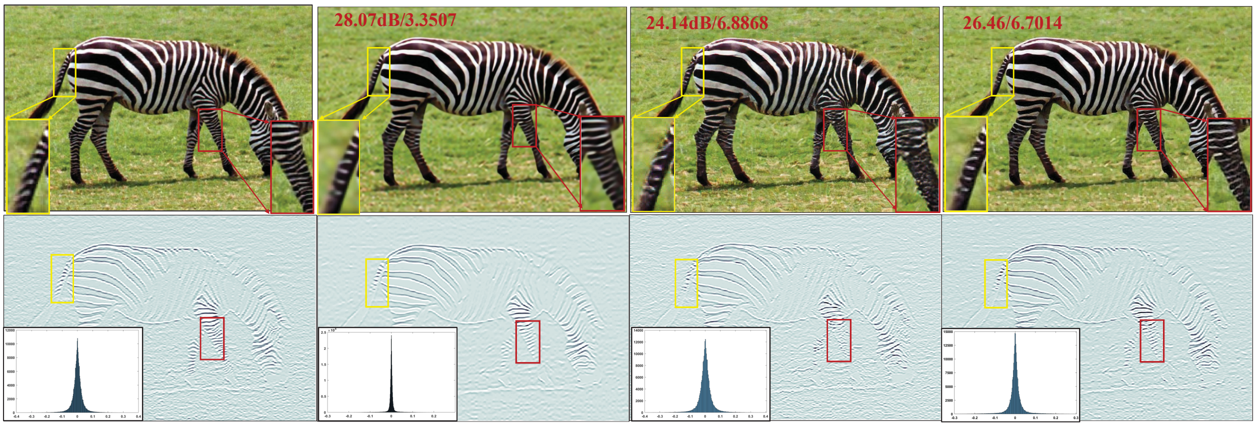
After normalization, for each high-frequency sub-band pair, the WDST algorithm is performed by minimizing a loss function that combines the content loss $L_c$, style loss $L_s$ [7] and a $\ell_1$ norm loss. The $\ell_1$ norm loss is specifically added to preserve the sparsity of wavelet coefficients. The total loss function for the $LH$ sub-band is defined as:

$$\mathcal{L}_{LH} = \alpha L_c(LH^r, LH^o) + \beta L_s(LH^r, LH^p) + \gamma \|LH^r\|_1, \tag{6}$$

where $\alpha$, $\beta$ and $\gamma$ are the weights for the content, style and $\ell_1$ norm loss, respectively. The content loss is defined as the MSE between the feature maps of the content input and the generated output at a specific layer $L$ of a pre-trained VGG network [23]:

$$L_c(LH^r, LH^o) = \frac{1}{2\sqrt{N_L M_L}} \sum_{i,j} (F^L_{ij}(LH^r) - F^L_{ij}(LH^o))^2. \tag{7}$$

Here, $F^L(LH^r)$ and $F^L(LH^o)$ are the feature maps at layer $L$ of a pre-trained VGG network [23] with $LH^r$ and $LH^o$ as inputs, respectively. In addition, $N_L$ is the number of feature maps at layer $L$, and $M_L$ is the product of the width and height of the feature map. Different from the content loss which is calculated between $LH^o$ and $LH^r$, the style loss is calculated between the style input $LH^p$ and $LH^r$. Moreover, unlike the content loss calculated at a single layer, the total style loss is defined by a weighed sum of




(a) GT (b) EDSR (content) (c) CX (style) (d) Ours

Figure 5. The first row shows the restored images of *Zebra* in Set 14 using EDSR, CX and our method, with the red values indicating the PSNR/NRQM values. The second row visualizes the HL sub-bands of the images in the first row, together with the histograms.

the style loss at different layers:

$$L_s(LH^r, LH^p) = \sum_l w_l L_s^l(LH^r, LH^p), \tag{8}$$

where $w_l$ is the weight for the style loss at the $l$-th layer. The $L_s^l(LH^r, LH^p)$ is calculated as the MSE between the Gram matrices of feature maps at the $l$-th layer in the pre-trained VGG network with $LH^r$ and $LH^p$ as inputs, respectively. Mathematically, it is defined as:

$$L_s^l(LH^r, LH^p) = \frac{1}{4N_l^2 M_l^2} \sum_{ij} (G_{ij}^l(LH^r) - G_{ij}^l(LH^p))^2, \tag{9}$$

where the $G^l(LH^r)$ and $G^l(LH^p)$ are the Gram matrices at the $l$-th layer for $LH^r$ and $LH^p$, respectively. We have $G^l(LH^r) = F^l(LH^r)^T F^l(LH^r)$ and $G^l(LH^p) = F^l(LH^p)^T F^l(LH^p)$. The layer conv 2-2 in VGG network [23] is used to calculate the content loss, and layers Relu1-1, Relu2-1, Relu3-1, Relu4-1, and Relu5-1 are used to calculate the style loss. With all loss defined, following [7], we use L-BFGS algorithm [32] to obtain $LH^r$ in (6) in a gradient decent way. Similarly, we can obtain $HL^r$ and $HH^r$.

After obtaining high-frequency sub-bands $LH^r$, $HL^r$, and $HH^r$, we need to de-normalize them. Then, we can reconstruct image $A_r$ by performing 2D ISWT on these high-frequency sub-bands together with the low-frequency sub-band $LL^r$ using the synthesis low-pass and high-pass filters $H_1$ and $G_1$. Here, for perfect reconstruction, $H_1$ and $G_1$ are the synthesis wavelet filters related to the analysis filters $H_0$ and $G_0$ used in the decomposition [17].

## 4. Numerical results

**Experimental setup.** For the 2D SWT, we use $bior2.2$ as the default wavelet filter. The number of wavelet decomposition levels is 2, which means we have six high-frequency sub-bands and one low-frequency sub-bands (see

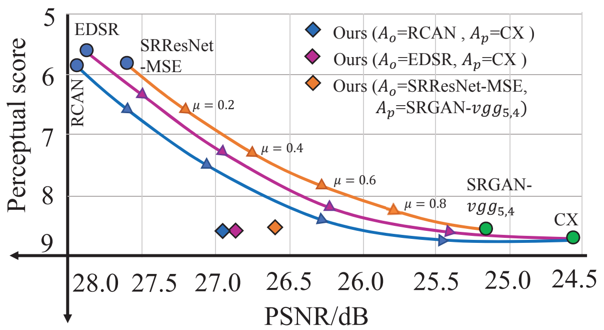
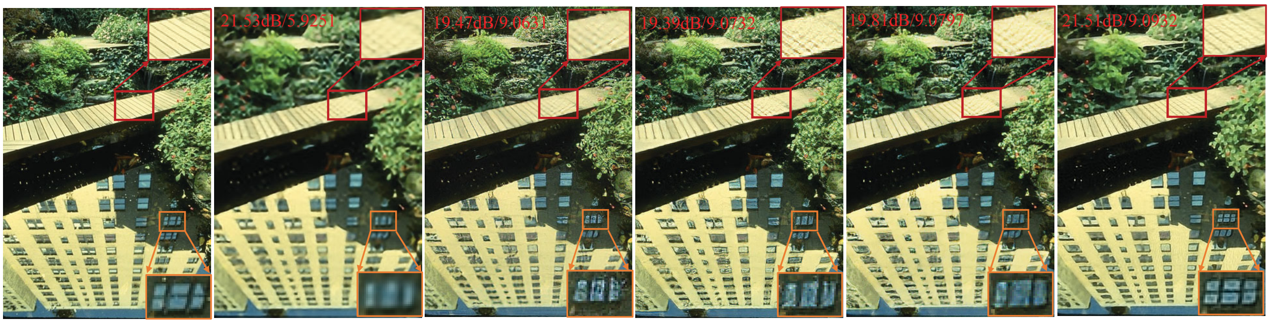



Figure 6. The perception-distortion (PD) curves of EDSR and CX, RCAN and CX, SRResNet-MSE and SRGAN-vgg54.

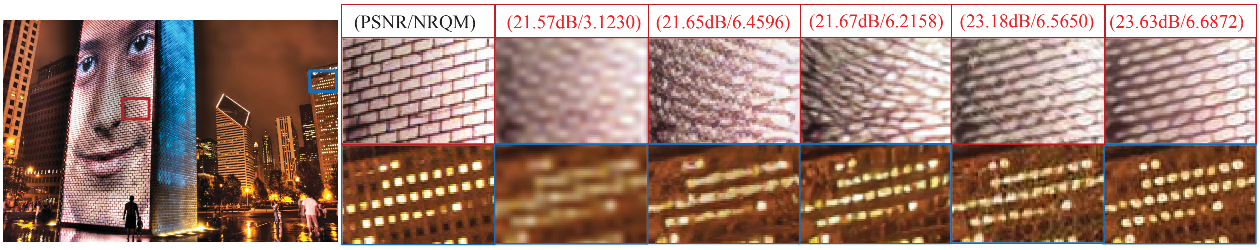
Fig. 2). In the LSE process, the loss function is minimized using the stochastic gradient descent (SGD) with backpropagation. The batch size is 64, the basic learning rate is 0.01 and the momentum is 0.9. In the WDST process, the ratio between the content loss and the style loss is $10^{-3}$, the ratio between the content loss and the $\ell_1$ norm loss is $10^{-5}$, and the weight of each layer when calculating the style loss is 0.2. The maximum iteration number is 5000 and 1000 for the first and second level decompositions, respectively. We use EDSR method [15] to obtain $A_o$, and CX method [19] to obtain $A_p$. Following [19], the perceptual score is calculated using NRQM [16]. We evaluate the performance of our method on various datasets, including Set5 [1], Set14 [29], BSD100 [18], Urban100 [9], and PIRM [2].

**Benchmarks.** The comparison methods are classified into three categories: methods that aim to improve the objective quality including A+ [25], Self-Ex [9], SRCNN [6], ESPCN [22], SRResNet-MSE [14], VDSR [12], EDSR [15], and RCAN [30]; methods that aim to improve the perceptual quality including SRGAN-vgg54 [14], SRGAN-vgg22[14], ENet [21], and CX [19]; and methods that aim to improve both the objective and perceptual quality



(a) GT (b) SRCNN (c) SRGAN_vgg54 (d) ENet (e) CX (f) Ours

Figure 7. Visual comparisons of image from BSD100 for 4× upscaling. The red numbers indicate the PSNR and NRQM values.




(a) GT (b) Bicubic (c) G-MGBP (d) ESRGAN (e) Deng (f) Ours

Figure 8. Visual comparisons of image from Urban100 for 4× upscaling. The red numbers indicate the PSNR and NRQM values.

Table 3. Effects of wavelet filter on Set 14 dataset.

| Filter | $haar$ | $db2$ | $bior2.2$ | $rbior2.2$ | $coif2$ | $db4$ | $bior4.4$ |
|---|---|---|---|---|---|---|---|
| PSNR | 28.06 | **28.08** | 28.07 | 27.96 | 28.05 | 28.06 | 28.05 |
| SSIM | **0.8379** | 0.8369 | 0.8356 | 0.8336 | 0.8344 | 0.8348 | 0.8343 |
| NRQM | 7.5109 | 7.6103 | 7.6827 | 7.6403 | 7.7101 | 7.6928 | **7.7442** |

Table 4. Ablation study of WDST on each sub-band.

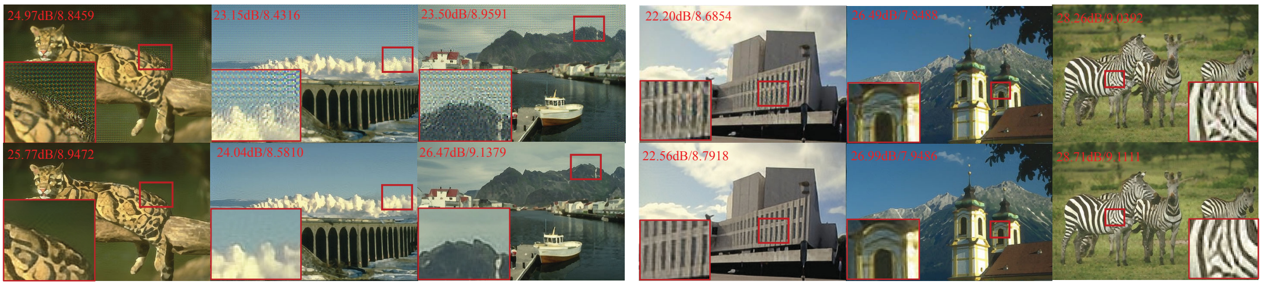
| Sub-band | $LH$ | $HL$ | $HH$ | PSNR | SSIM | NRQM |
|---|---|---|---|---|---|---|
| WDST | N | Y | Y | 27.19 | 0.7195 | 7.8490 |
| | Y | N | Y | 27.28 | 0.7227 | 7.8343 |
| | Y | Y | N | 26.96 | 0.7105 | 8.0542 |
| | Y | Y | Y | 26.82 | 0.7058 | 8.5948 |

including SRGAN-MSE [14], G–MGBP[20], PESR [26], EUSR[4], Deng [5] and ESRGAN[27].

**Effectiveness of WDST.** In order to show the effectiveness of our WDST algorithm, we visualize in Fig. 5 the input content and style sub-bands, as well as the output sub-band using the WDST algorithm. As can be seen, the content sub-band lacks many high-frequency details and the style sub-band has messy structures, e.g, the horse leg and tail. After the WDST, the output sub-band overcomes these drawbacks, which is now abundant in high-frequency details and has clear textures and structures. To some extent, the output sub-band corrects the wrong information in the style sub-band and re-locate it in the right place, with the guidance of the content sub-band. We also show the histogram distributions of the sub-bands in Fig. 5. It can be seen that our histogram is closer to the ground-truth compared to EDSR, which is the reason why we have higher perceptual quality.

**Wavelet filter sensitivity.** In our algorithm, we use wavelet filter to decompose each image into various sub-bands. In order to investigate the effects of wavelet filter on the performance of our algorithm, we present in Table 3 the PSNR, SSIM and NRQM results with different wavelet filters. These filters include $haar$, $db2$ and $db4$ from Daubechies, $bior2.2$ and $bior4.4$ from Biorthogonal, $rbio2.2$ from Reverse biorthogonal, and $coif2$ from Coifman wavelet family. From Table 3, we can see that the wavelet filter indeed has some effects on the performance. Specifically, the $haar$ filter has the highest SSIM value, the $db2$ filter performs best in PSNR and the $bior4.4$ filter has the best perceptual quality. However, the difference among different filters is not very significant.

**Perception-distortion (PD) performance.** Fig. 1 compares the PD performance of different methods in the PSNR and NRQM plane. As we can see, methods A+, Self-Ex, SRCNN, ESPCN, SRResNet-MSE, VDSR, EDSR, RCAN occupy the upper left region which means they have high objective quality but low perceptual quality. In contrast, methods SRGAN-vgg54, SRGAN-vgg22, ENet, and CX take up the bottom right region, which indicates they have high perceptual quality but low objective quality. Other methods like SRGAN-MSE, PESR, Deng, and ESRGAN stand in the middle region, which are all trying to achieve a good tradeoff between distortion and perceptual quality. Among all these methods, our method is the closest to the




(a) SRGAN-MSE [14] and ours

(b) Deng [5] and ours

Figure 9. (a) compares the images between SRGAN-MSE and ours, (b) compares the images of Deng and ours. The first rows in (a) and (b) are SRGAN-MSE and Deng, and the second row is our method. The red numbers indicate the PSNR and NRQM values.

bottom left corner, which means that we achieve the best trade-off between the objective and perceptual quality. Table 2 compares the numerical results of our method with SRGAN-MSE [14], G–MGBP[20] , PESR [26], Deng [5] and ESRGAN[27] (with $\alpha$ =0.8), which all aim to improve both the perceptual and objective quality. As we can see, our method outperforms others in both perceptual and objective quality.

**Content and Style inputs sensitivity.** To show the position of our method more clearly, we draw in Fig. 6 the PD curve of EDSR and CX, which are the two default methods to generate $A_o$ and $A_p$ in this paper. The curve is drawn by interpolating the pixel values of $A_o$ and $A_p$ with a parameter $\mu \in [0, 1]$, as follows

$$A_r = \mu * A_p + (1 - \mu) * A_o. \tag{10}$$

Obviously, when $\mu$ increases, the NRQM increases while the PSNR decreases. As we can see from Fig. 6, our method is far lower than that PD curve, which means we are much better than the simple interpolation of $A_o$ and $A_p$. To investigate our sensitivity to the content and style inputs, we also draw the PD curves of RCAN [30] and CX, SRResNet-MSE and SRGAN-vgg54 [14], together with our correspoding results. We can see that, even in the worst case (with SRResNet-MSE and SRGAN-vgg54 as inputs), our algorithm still achieves better PD trade-off (i.e., PSNR/NRQM=26.56 dB/8.5005) than Deng (26.46 dB/8.4452) and ESRGAN (26.44 dB/8.3034).

**Visual comparison.** Figs. 7 and 8 visualize the images of our and other methods. We can see from Fig. 7 that our method can restore correctly the texture of the bridge and the structure of the window, while others either distort the texture or struggle to restore the structure. From Fig. 8, we can see that our method can restore the wall and lights clearly, while others fail to do so. Our method also overcomes many drawbacks of other methods. Fig. 9 (a) compares our method with SRGAN-MSE [14]. We can see that the SRGAN-MSE method produces lots of abnormal noise and wrong textures in the images, while our method does not have these problems. Fig. 9 (b) compares our method with Deng [5], which shows that the images of Deng [5] are noisy and have messy structures. In contrast, our method is able to reconstruct images with clean and accurate structures.

**Ablation study.** In order to study the effects of each high-frequency sub-band on the perception-distortion performance, we show in Table 4 the results when WDST is not performed on one of the sub-bands. From this table, we can see that each sub-band contributes to the perception-distortion performance. When WDST is absent from any of them, the perceptual quality (NRQM) decreases significantly. However, compared with LH and HL sub-bands, the influence of HH sub-band is not very significant. This is because the HH sub-band contains the diagnonal information, which is not as much as the horizontal and vertical information contained in the LH and HL sub-bands, respectively.

## 5. Conclusion and future work

In this paper, we have proposed a novel method based on wavelet domain style transfer, to give an excellent solution to the perception-distortion conflict in SISR. We find that the objective and perceptual quality are influenced by different elements of an image. To achieve the best trade-off between them, we use stationary wavelet decomposition (SWT) to split elements related with objective quality from those related with perceptual quality. Then, we can optimize each with different targets, with little influence on the other. This "divide and conquer" strategy was demonstrated to achieve a good trade-off between the image distortion and perception, and we believe this can inspire more follow-up works to further push forward the reconstruction performance in SISR. Like the conventional style transfer work [14], we need many iterations to solve the optimization problem in (6), which is a little bit time-consuming, i.e., around 60 seconds for each sub-band. Inspired by the real-time artistic style transfer work [11], our future work is to train a feed-forward network to predict the fused sub-band which minimizes (6), so that the computational complexity can be significantly decreased.

## References


[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British machine vision conference (BMVC)*, pages 1–10. IEEE, 2012.

[2] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 PIRM challenge on perceptual image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 334–355. Springer, 2018.

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6228–6237, 2018.

[4] Jun-Ho Choi, Jun-Hyuk Kim, Manri Cheon, and Jong-Seok Lee. Deep learning-based image super-resolution considering quantitative and perceptual quality. *arXiv preprint arXiv:1809.04789*, 2018.

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

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

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016.

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

[9] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015.

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

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016.

[12] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016.

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, 2016.

[14] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.

[15] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, volume 1, page 3, 2017.

[16] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

[17] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):674–693, 1989.

[18] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 416–423. IEEE, 2001.

[19] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Maintaining natural image statistics with the contextual loss. In *Asian Conference on Computer Vision*, pages 427–443. Springer, 2018.

[20] Pablo Navarrete Michelini, Dan Zhu, and Hanwen Liu. Multi–scale recursive and perception–distortion controllable image super–resolution. In *European Conference on Computer Vision (ECCV)*, pages 3–19. Springer, 2018.

[21] Mehdi SM Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *International Conference on Computer Vision (ICCV)*, pages 4501–4510. IEEE, 2017.

[22] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

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

[24] Jean-Luc Starck, Jalal Fadili, and Fionn Murtagh. The undecimated wavelet decomposition and its reconstruction. *IEEE Transactions on Image Processing*, 16(2):297–309, 2007.

[25] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision (ACCV)*, pages 111–126. Springer, 2014.

[26] Thang Vu, Tung M Luu, and Chang D Yoo. Perception-enhanced image super-resolution via relativistic generative adversarial networks. In *European Conference on Computer Vision*, pages 98–113. Springer, 2018.

[27] 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 *European Conference on Computer Vision (ECCV)*, pages 63–79. Springer, 2018.

[28] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.

[29] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pages 711–730. Springer, 2010.

[30] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[31] Zhisheng Zhong, Tiancheng Shen, Yibo Yang, Zhouchen Lin, and Chao Zhang. Joint sub-bands learning with clique structures for wavelet domain super-resolution. In *Advances in Neural Information Processing Systems*, pages 165–175, 2018.

[32] Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.