



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Ph.D. DISSERTATION

Deep Learning of Perception-Oriented  
Image Restoration using Conditional  
Objective

조건 목적을 사용한 딥러닝 기반의 인지적 영상 복원

BY

Seung Ho Park

AUGUST 2023

DEPARTMENT OF ELECTRICAL AND  
COMPUTER ENGINEERING  
COLLEGE OF ENGINEERING  
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Deep Learning of Perception-Oriented  
Image Restoration using Conditional  
Objective

조건 목적을 사용한 딥러닝 기반의 인지적 영상 복원

BY

Seung Ho Park

AUGUST 2023

DEPARTMENT OF ELECTRICAL AND  
COMPUTER ENGINEERING  
COLLEGE OF ENGINEERING  
SEOUL NATIONAL UNIVERSITY

# Deep Learning of Perception-Oriented Image Restoration using Conditional Objective

조건 목적을 사용한 딥러닝 기반의 인지적 영상 복원

지도교수 조 남 익  
이 논문을 공학박사 학위논문으로 제출함

2023년 8월

서울대학교 대학원

전기정보공학부

박승호

박승호의 공학박사 학위 논문을 인준함

2023년 8월

위 원 장	이 종 호
부 위 원 장	조 남 익
위 원	전 세 영
위 원	김 영 민
위 원	구 형 일

# Abstract

The purpose of image restoration (IR) is to reconstruct a high-quality (HQ) image corresponding to a given low-quality (LQ) image. Typical image restoration tasks include image denoising and image super-resolution. IR has many applications, mainly as a pre-processing step of image enhancement, computer vision, or image analysis tasks, such as medical, surveillance, and satellite image analysis. However, it is challenging since IR is an ill-posed problem in that infinitely many HQ images correspond to a single LQ image. Recently, the performance of IR has been greatly improved by adopting deep neural networks trained with large-scale external datasets. Pixel-wise distortion-oriented losses (L1 and L2) were widely used in early research, which helped to obtain a high signal-to-noise ratio (PSNR). However, these losses lead the model to generate an average of possible HQ solutions, which are usually blurry and thus visually not pleasing. Subsequently, perception-oriented losses, such as perceptual loss and generative adversarial loss, were introduced to overcome this problem and produce realistic images with fine details. Although these perception-oriented losses are used for various IR methods, they also bring undesirable side effects, such as unnatural details and structural distortions. It has been shown that using a single perceptual loss is insufficient for accurately restoring locally varying diverse shapes in images. For this reason, combinations of various losses, such as perceptual, adversarial, and distortion losses, have been attempted, yet it remains challenging to find optimal combinations. To address these problems, this dissertation presents a new method that applies desired or optimal objectives for each region to generate plausible results in overall areas of high-quality outputs. This dissertation first proposes an efficient learning method that enables a single super-resolution (SR) model to produce reconstruction results in a locally flexible style. A typical approach to obtaining alternative SR results is to train multiple SR models with different loss weightings and exploit the combi-

nation of these models. Instead of using multiple models, I propose a method to optimize an SR model with a conditional objective during training, where the objective is a weighted sum of multiple perceptual losses at different feature levels. The weights vary according to given conditions, and the set of weights is defined as a style controller. Also, I present an architecture appropriate for this training scheme: the Residual-in-Residual Dense Block equipped with spatial feature transformation layers. The trained model can generate locally different outputs conditioned on the style control map at the inference phase. Extensive experiments show that the proposed SR model produces various desirable reconstructions without artifacts and yields comparable quantitative performance to state-of-the-art SR methods. Second, this dissertation also presents a new SR framework for perception-oriented restoration by estimating locally optimal objectives for each region to generate plausible results in overall areas of high-quality outputs. Specifically, the framework consists of two models: a predictive model that infers an optimal objective map for a given low-resolution (LR) input and a generative model that applies a target objective map to produce the corresponding SR output. The generative model is trained over the proposed objective trajectory representing a set of essential objectives, which enables the single network to learn various SR results corresponding to combined losses on the trajectory. The predictive model is trained using pairs of LR images and corresponding optimal objective maps searched from the objective trajectory. Experimental results on five benchmarks show that the proposed method outperforms state-of-the-art perception-driven SR methods in LPIPS, DISTs, PSNR, and SSIM metrics. The visual results also demonstrate the superiority of the proposed method in perception-oriented reconstruction.

**keywords:** image restoration, image super-resolution, perception-oriented image restoration, perception-oriented image super-resolution, conditional objective, optimal objective estimation

**student number:** 2015-31015

# Contents

<b>Abstract</b>	<b>i</b>
<b>Contents</b>	<b>iii</b>
<b>List of Tables</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Contribution . . . . .	4
1.2 Contents . . . . .	5
<b>2 Flexible Style Image Super-Resolution using Conditional Objective</b>	<b>6</b>
2.1 Motivation and Overview . . . . .	6
2.2 Related Work . . . . .	8
2.2.1 Loss Functions for SISR . . . . .	8
2.2.2 Network Conditioning . . . . .	9
2.2.3 Continuous Imagery Effect Transition . . . . .	10
2.2.4 Multi-task Learning . . . . .	10
2.3 Proposed Method . . . . .	11
2.3.1 Targeted Perceptual Loss . . . . .	11
2.3.2 Proposed SR with Flexible Style . . . . .	12
2.3.3 Proposed Network Architecture . . . . .	14

2.3.4	Proposed Loss Function . . . . .	16
2.3.5	Implementation details . . . . .	17
2.4	Experiments . . . . .	19
2.4.1	Materials and Methods . . . . .	19
2.4.2	Evaluation of Flexible SR for Perception-Distortion (FxSR-PD)	23
2.4.3	Flexible SR for Diverse Styles (FxSR-DS) . . . . .	27
2.4.4	Per-pixel Style Control . . . . .	29
2.4.5	Compressed LR Image Restoration . . . . .	31
2.4.6	Complexity Analysis . . . . .	31
2.4.7	Ablation Study . . . . .	32
2.4.8	Discussion . . . . .	33
2.5	Conclusion . . . . .	34
<b>3</b>	<b>Perception-Oriented Single Image Super-Resolution using Optimal Ob- jective Estimation</b>	<b>41</b>
3.1	Motivation and Overview . . . . .	41
3.2	Related Work . . . . .	45
3.3	Methods . . . . .	46
3.3.1	Proposed SISR Framework . . . . .	46
3.3.2	Proposed Generative Model . . . . .	47
3.3.3	Optimal Objective Estimation (OOE) . . . . .	54
3.4	Experiments . . . . .	57
3.4.1	Experiment Setup . . . . .	57
3.4.2	Evaluation . . . . .	57
3.5	Ablation Study . . . . .	58
3.6	Conclusion . . . . .	59
<b>4</b>	<b>Conclusions</b>	<b>63</b>
	<b>Bibliography</b>	<b>65</b>





# List of Tables

2.1	Comparison with state-of-the-art SR methods on benchmarks. In $4\times$ and $8\times$ , the 1st and the 2nd best performances are highlighted in <b>bold</b> and <u>underline</u> , respectively. . . . .	22
2.2	Comparison with state-of-the-art SR methods on DIV2K in terms of low resolution consistency, photo-realism and meaningful diversity. The numbers in the parentheses are the relative performances, i.e., the best value is set to 1, and the others are divided by the best value. . . .	28
2.3	The details of video we used for SR performance comparison. . . . .	31
2.4	Comparisons of the running time, the computational costs and the size of SR models for super-resolution $4\times$ , when the size of LR input images is $128 \times 128$ . . . . .	32
3.1	Performance comparison of SR results of ESRGAN models with different weight vectors for perceptual loss. Among the objectives in Sets $A$ and $B$ , except for $\lambda_0$ , the 1st and the 2nd best performances for each column are highlighted in <b>bold</b> and <u>underline</u> . . . . .	60
3.2	Comparison with state-of-the-art SR methods on benchmarks. The 1st and the 2nd best performances for each group are highlighted in <b>bold</b> and <u>underline</u> , respectively. LR-PSNR values greater than 45dB are written in <i>italic</i> . . . . .	61

3.3 Comparison of performance according to different selections. The bold checkmark indicates the change from the left selection. The 1st and the 2nd best performances except for SROOS are highlighted in **bold** and underline, respectively. . . . . 62

3.4 Comparison of the running time and the SR model size. . . . . 62

# List of Figures

1.1	(a) A high-resolution image, (b) the degraded image with downscale factor 4, (c) the realistic restoration result of perception-oriented SR $4\times$ , (d) the blurry restoration result of distortion-oriented SR $4\times$ . . . . .	3
1.2	$4\times$ SR performance comparison. (a) whole image, (b) high resolution image, (c) the restoration results of the conventional distortion-oriented SR, (d) the restoration results of the conventional perception-oriented SR, (e) the restoration results of the proposed SR . . . . .	4
1.3	The perception-distortion tradeoff. Image restoration algorithms can be characterized by their average distortion and by the perceptual quality of the images they produce. . . . .	5
2.1	The effect of choosing different layers when estimating perceptual losses on different regions, e.g., on edge and texture regions, where the losses correspond to MSE, ReLU 2-2 (VGG22), and ReLU 4-4 (VGG44) of the VGG-19 network. . . . .	8
2.2	The architecture of our proposed flexible SR network. We use the RRDB equipped with SFT as a basic block (Figure ??(c)). The condition branch takes a style map for reconstruction style as input. This map is used to control the recovery styles of edges and textures for each region through SFT layers. . . . .	11
2.3	RRDB with SFT for basic blocks . . . . .	13

2.4	The left column shows the weight functions for FxSR-PD. $t = 0$ corresponds to distortion-oriented SR (only MSE loss) and $t = 1$ perception-oriented (with adversarial and perceptual loss from VGG22). The right column shows the weight functions for FxSR-DS, where more perceptual losses are used to expand the HR styles. . . . .	15
2.5	Changes in the result of FxSR-PD $4\times$ SR according to $t$ on DIV2K validation set [?]. . . . .	18
2.6	$4\times$ SR performance comparison of state-of-the-art and proposed methods evaluated by the metric (a) PSNR, (b) SSIM [?], and (c) MS-SSIM [?] for DIV2K according to condition parameters. . . . .	19
2.7	$4\times$ SR performance comparison of state-of-the-art and proposed methods evaluated by the metric (a) LPIPS [?], (b) DISTS [?], and (c) VIF [?] for DIV2K according to condition parameters. . . . .	20
2.8	$4\times$ SR performance comparison of state-of-the-art and proposed methods evaluated by the (a) NIQE [?] and (b) BRISQUE [?] for DIV2K according to condition parameters. . . . .	20
2.9	Performance comparison of the state-of-the-arts and proposed method (FxSR-PD: Perception-Distortion Flexible SR) for DIV2K $4\times$ SR. . .	21
2.10	$4\times$ SR P-D performance comparison of state-of-the-art and proposed methods evaluated by the metric (a) NIQE [?] Vs. PSNR, (b) SSIM [?] Vs. LPIPS [?], and (c) SSIM [?] Vs. DISTS [?] for DIV2K according to condition parameters. . . . .	23
2.11	Visual comparison with state-of-the-art perception-driven SR methods on DIV2K validation set [?]. The proposed method produces competitive results compared to other modern techniques and can also generate reconstructed images of various styles of LR images. . . . .	24
2.12	Visual comparison for $8\times$ SR results on DIV2K validation set [?]. . .	25

2.13	Changes in the result of FxSR-DS $4\times$ SR according to $t$ on DIV2K validation set [?]. . . . .	26
2.14	On the left are the SR results of FxSR-PD (top) and FxSR-DS (bottom) for DIV2K 0858, corresponding to $t$ values with Global Best (G-Best) LPIPS among 11 samples, respectively. In the middle are the LPIPS maps of the SR results on the left. On the right are the Local Best (L-Best) LPIPS maps generated by selecting the highest score per pixel from 11 samples. The brighter the pixel, the higher the LPIPS value and the greater the perceptual difference from the ground truth. Each number in parentheses is the average LPIPS value for the entire image.	35
2.15	Comparison of the SR results of the conventional method (a), which applies one objective to the entire image, and the FxSR-PD method, which applies different objectives for each area (clothes and letters) through a local map. We can see that the proposed FxSR-PD in (b) can more accurately produce the locally intended and suitable SR results without side effects such as blurry textures and broken characters. . . .	36
2.16	Comparison of the SR results of the conventional method (a), which applies one objective to the entire image, and the FxSR-DS method, which applies different objectives for each area (buildings and trees) through a local map. We can see that the proposed FxSR-DS in (b) can more accurately produce the locally intended and suitable SR results without side effects such as blurry tree textures and overshoot around the edges. . . . .	37
2.17	Depth-adaptive FxSR. <b>T</b> -maps is the modified version of the depth map of an image from the Make3D dataset [?] . . . . .	38
2.18	An example of applying a user-created depth map to enhance the perspective feeling with the sharper and richer textured foreground and the background with more reduced camera noise than the ground truth.	38

2.19	The SR Results for compressed LR images. Two feature space (VGG44 and VGG54) and 16 RBs with SFT are used for FxSR-CA model. LR Images are extracted from "Amazing Place" video title that is encoded by VP9 codec at 0.3Mbps. . . . .	39
2.20	Convergence of diversity curve of the proposed FxSR-PD model as the number of training iteration increase, using (a) 16 RBs with SFT and (b) using 23 RRDBs with SFT. (c) The performance comparison between two FxSR-PD version at the 250,000th iteration . . . . .	40
3.1	Visual and quantitative comparison. The proposed SROOE shows a higher PSNR, LR-PSNR [?] and lower LPIPS [?] than other state-of-the-art methods, <i>i.e.</i> , , lower distortion and higher perceptual quality. .	42
3.2	Architecture of the proposed method. The predictive model generates the optimal objective map $\hat{\mathbf{T}}_B$ , which is fed to the generative model. The input LR image is super-resolved through our Basic Blocks and other elements of the generator, which are controlled by the map from the Condition Branch. . . . .	46
3.3	Set $A$ (left) and set $B$ (right) in the objective space. The objectives in set $B$ are closer each other than those in set $A$ . . . . .	48
3.4	The $OOS_A$ and $OOS_B$ results using Sets $A$ and $B$ (top), their SR results, ESRGAN- $OOS_A$ and ESRGAN- $OOS_B$ (bottom). . . . .	49
3.5	The proposed vector functions for loss weights, (a) $\lambda(t)$ in Eqn. ?? when $\alpha=1$ and $\beta=0$ , (b) its $\lambda_{per}(t)$ and (c) the weighting functions for $\lambda_{per}(t)$ . (d) $\lambda_{per}(t)$ used for FxSR [?]. . . . .	51
3.6	Changes in detail in the SROT results according to $t$ -value (top) and changes in PSNR and LPIPS for test DBs (bottom). . . . .	53
3.7	The input image, the optimal objective selection $\mathbf{T}_S^*$ obtained by parameter sweeping, and $\hat{\mathbf{T}}_B$ estimated by $C_\psi$ . . . . .	55

3.8	Visual comparison with state-of-the-art SR methods. Among the seven perception-oriented SR methods, the best performances are highlighted in <b>bold</b> . . . . .	56
3.9	Visual comparison of the results of FxSR and SROOE. . . . .	58



# Chapter 1

## Introduction

The widespread adoption of smartphones with cameras has made it easier for people to capture images and record videos. Smartphones have become a primary tool for content creation and consumption. The popularity of social media and content-sharing platforms has led to a surge in image and video consumption. These platforms attract billions of views every day. However, in many real-world scenarios, images are captured in low-light conditions or with noisy sensors, resulting in poor image quality. Image restoration (IR) aims to reconstruct a high-quality (HQ) image corresponding to a given degraded low-quality (LQ) image. Noise in an image refers to random variations in pixel values that may arise during the image acquisition or transmission process, and denoising helps in removing the noise, leading to clearer and more visually appealing images. Similarly, super-resolution enhances the resolution of low-resolution images, making them more detailed and informative. By enhancing the details and sharpness of an image, it becomes more visually appealing and easier to interpret. Therefore, they are essential preprocessing steps in various applications like photography, video streaming, medical imaging, satellite imagery, and more. In addition, as many computer vision tasks, such as object detection, image segmentation, and recognition, heavily rely on the quality of input images, the performance of these downstream tasks can significantly improve with these image restoration techniques.

In the image restoration framework, an LQ image  $y$  is usually modeled as one of the outputs of the following degradation processes:

$$\mathbf{y} = T((\mathbf{x} \otimes k) \downarrow_s) \quad (1.1)$$

$$\mathbf{y} = (\mathbf{x} \otimes k) \downarrow_s + \mathbf{n} \quad (1.2)$$

where  $\mathbf{x} \otimes k$  represents the convolution between a blur kernel  $k$  and a HR image  $x$ ,  $\downarrow_s$  is a subsequent downsampling operation with scale factor  $s$ ,  $T$  and  $n$  are compression operation and additive white Gaussian noise (AWGN), respectively.

Recently, the performance of image restoration has been greatly improved by adopting deep neural networks trained with a large amount of image data. However, distortion-oriented losses such as  $L1$  or  $L2$  lead the model to produce an average of possible HQ solutions, resulting in a blurry, visually not pleasing image. Subsequently, perception-oriented losses were introduced to overcome this problem and to generate high-contrast results. Fig. ?? compares the resolution restoration results between distortion-driven and perception-driven SR. It can be seen that the perception-driven result is more perceptually similar to the ground truth than the distortion-oriented result, and it is also more visually pleasing.

Although these perception-oriented losses are used for various IR methods [?, ?, ?], they also bring undesirable side effects such as unnatural details and structural distortions. Fig. ?? shows some examples of the sharp but side-effected results of the perception-oriented SR compared to the blurry results of the distortion-oriented SR. Blau [?] argued that it is difficult to simultaneously achieve perceptual quality enhancement and distortion reduction because they involve a trade-off relationship as shown in Fig. ?. However, observing Fig. ??, it can be seen that when the perception and distortion-oriented SR results are appropriately applied differently for each region according to the context, high contrast and realistic SR results with much-reduced artifacts can be achieved. One of the main claims of this dissertation is that the proposed method can further reduce distortion and increase perceptual quality simultaneously,

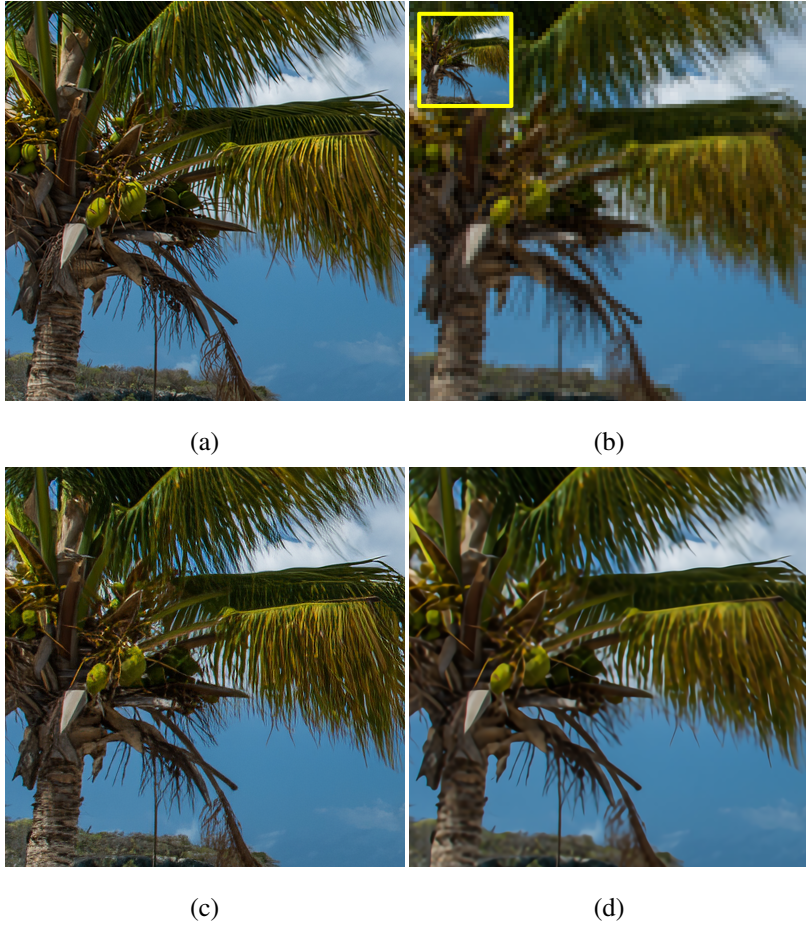


Figure 1.1: (a) A high-resolution image, (b) the degraded image with downscale factor 4, (c) the realistic restoration result of perception-oriented SR  $4\times$ , (d) the blurry restoration result of distortion-oriented SR  $4\times$ .

as the blue solid line in Fig. ?? is shifted to the red dotted line. The results of the proposed method adaptively applying distortion-oriented and perceptual-oriented results to each region are shown in the last column of Fig. ?. In this dissertation, it is also verified that the proposed perception-oriented flexible SR works well for compressed JPEG images and compressed streaming videos.

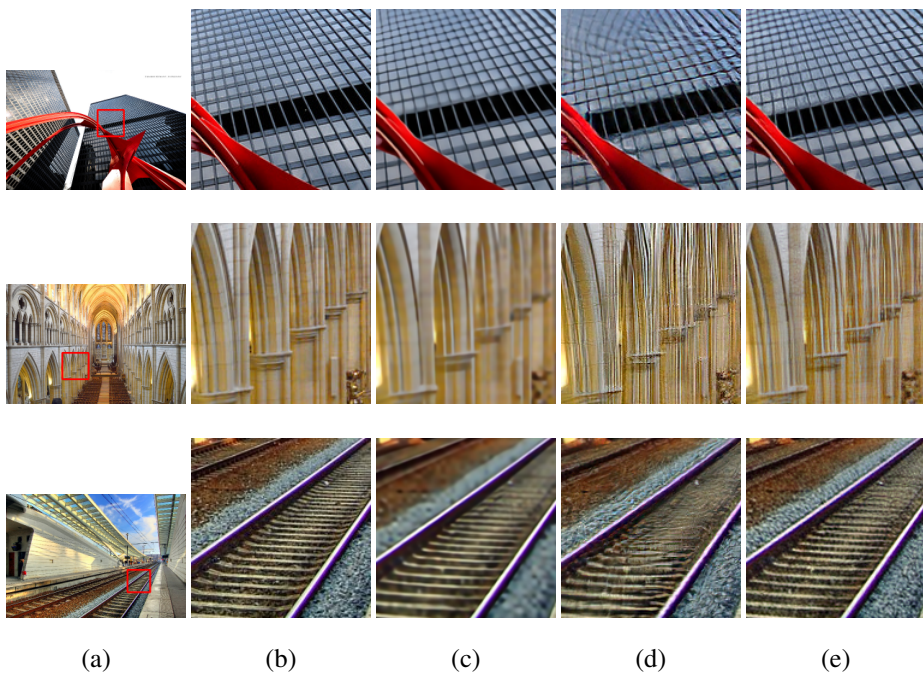


Figure 1.2:  $4\times$  SR performance comparison. (a) whole image, (b) high resolution image, (c) the restoration results of the conventional distortion-oriented SR, (d) the restoration results of the conventional perception-oriented SR, (e) the restoration results of the proposed SR

## 1.1 Contribution

Our contributions are summarized as follows. (1) This dissertation presents an efficient method to train a single locally-adjustable model for perception-oriented restoration by using a conditional objective. (2) This dissertation also proposes a new framework for perception-oriented restoration that estimates and applies an optimal combination of objectives for each input region and thus produces perceptually accurate restoration results. (3) The proposed methods are validated for low-resolution compressed images and videos. (4) The proposed methods achieve state-of-the-art performance.

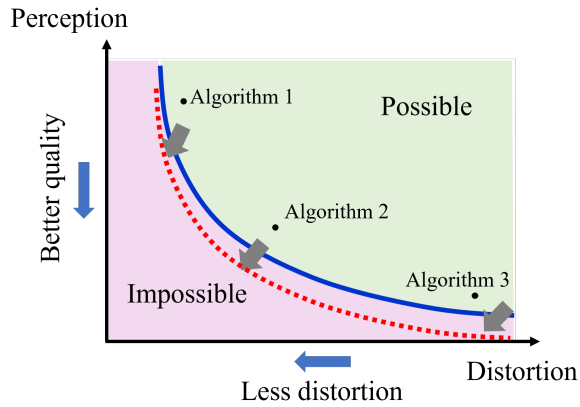


Figure 1.3: The perception-distortion tradeoff. Image restoration algorithms can be characterized by their average distortion and by the perceptual quality of the images they produce.

## 1.2 Contents

The rest of this dissertation is organized as follows. In Chapter 2, flexible style single image super-resolution using conditional objective is proposed. Chapter 3 introduces perception-oriented single image super-resolution using optimal objective estimation. Finally, this dissertation is concluded in chapter 4.

## Chapter 2

# Flexible Style Image Super-Resolution using Conditional Objective

### 2.1 Motivation and Overview

Finding a high-resolution (HR) counterpart from a given low-resolution (LR) image is referred to as single image super-resolution (SISR). The SISR is an ill-posed problem in that infinitely many HR images correspond to a single LR image. Despite such ill-posedness, recent convolutional neural networks (CNNs) are shown to map an LR to a plausible HR [?].

SRCNN [?, ?] first showed the effectiveness of a CNN for SISR, and various CNN architectures have been proposed for better performance afterward [?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?]. Earlier works used mean square error (MSE) as a loss function to train the network. However, since it tends to produce blurry HR outputs, researchers are finding new loss functions to generate more realistic outputs [?, ?]. Specifically, perceptual losses [?] are introduced to optimize the super-resolution (SR) model in the feature space instead of pixel space. Ledig *et al.* [?] proposed to use adversarial loss [?] in combination with the perceptual loss to encourage the network to favor perceptually superior solutions residing in the manifold of natural images.

More recently, Wang *et al.* [?] investigated class-conditional SR. It employed Spatial Feature Transform (SFT) capable of altering an SR network’s behavior conditioned on semantic segmentation probability maps. However, since most of the existing methods calculate perceptual losses on an entire image in the same feature space, the results tend to be monotonous and unnatural. For this reason, Rad *et al.* [?] optimized SR models with a targeted objective function that penalizes images at different semantics using the corresponding terms. But, since the segmentation label needs to be fed to the SR network to calculate the targeted perceptual loss, the users cannot easily adjust the objective function. In summary, most early SR networks provide a designated HR output among many possible ones, not allowing us to explore more plausible outputs at the test phase. To alleviate this problem, Lugmayr *et al.* [?] proposed the SRFlow using a normalizing flow method capable of learning the conditional distribution of the output given the low-resolution input. As a result, it can learn to predict diverse photo-realistic high-resolution images.

Though great strides have been made, the natural and flexible reconstruction of local regions is still challenging. As stated previously, there can be diverse HR solutions for a given LR, meaning that one LR input can be restored to different HR results depending on the context and situation. Particularly because of various shapes and textures in the real world, the one-to-many problem becomes even more serious if the SR network’s capacity is not large enough.

To solve this problem, first, the SR model should be able to generate more diverse styles of HR reconstruction while keeping consistency with the given LR image. Second, the recovery style needs to be locally controlled. Third, training and storing too many redundant SR models with different parameters should be avoided. Achieving these requirements would enable us to explore various HR solutions for each region effectively. In this respect, some recent methods made it possible to continuously generate and adjust intermediate results between two objective functions, *i.e.*, perception and distortion functions [?, ?, ?]. However, there can be some improvements in these

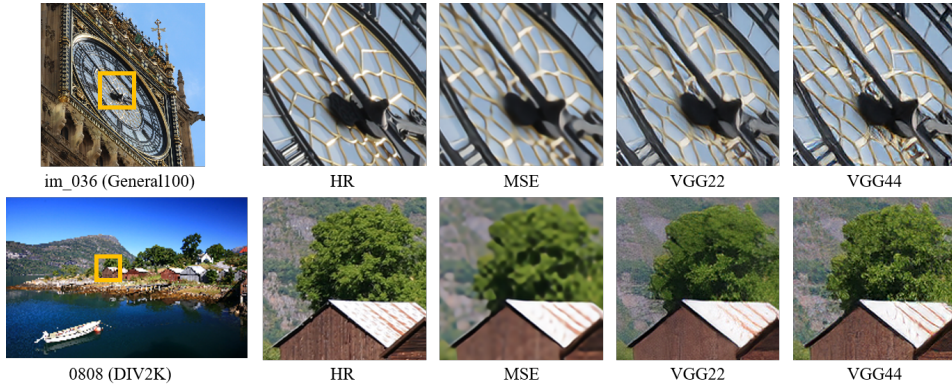


Figure 2.1: The effect of choosing different layers when estimating perceptual losses on different regions, e.g., on edge and texture regions, where the losses correspond to MSE, ReLU 2-2 (VGG22), and ReLU 4-4 (VGG44) of the VGG-19 network.

approaches, as they defined just two objective functions and controlled the entire image, not the local regions needing adjustment.

In this paper, we attempt locally adjustable HR generation by exploring the SR model optimization, focusing on the development of conditional objectives that can generate various reconstruction styles. The proposed objective consists of the weighted sum of several perceptual losses from different feature levels. The weights vary according to the condition, which is the recovery style information in our work. Experiments show that training an SR model with our multi-level perceptual losses generates various recovery styles effectively, which also enables us to finely control the styles of local regions.

## 2.2 Related Work

### 2.2.1 Loss Functions for SISR

The choice of the objective function affects the recovery style and reconstruction performance. For instance, adversarial loss [?] encourages an SR network to generate



perception-oriented solutions [?, ?, ?, ?]. Perceptual losses [?, ?] are proposed to optimize SR models by minimizing the error in the feature space instead of pixel space. Dovovitskiy *et al.* [?] and Ledig *et al.* [?] proposed to use adversarial loss in combination with the perceptual loss to encourage the network to favor solutions that look more like natural images. With these loss functions, the overall visual quality of reconstruction is significantly improved [?, ?, ?]. Recently, some studies [?, ?, ?] proposed to use GAN with losses based on perceptual quality assessment metric. Another perceptual loss is proposed in [?], using different levels of features according to semantic segmentation labels such as objects, boundaries, and backgrounds. In these approaches, once an SR model is trained, a fixed HR is produced for the LR input.

### 2.2.2 Network Conditioning

The feature normalization techniques generally change networks' behavior based on the input properties. The representative normalization methods may be batch normalization (BN) [?] and instance normalization (IN) [?]. The IN normalizes a single image while the BN does a whole batch of images. Conditional Instance Normalization (CIN) has also been introduced in [?], which uses the learned representations to model multiple styles simultaneously. Huang *et al.* [?] proposed adaptive instance normalization (AdaIN) to adjust features to arbitrary new styles. Perez *et al.* [?] proposed Feature-wise Linear Modulation, called FiLM, as a general-purpose conditioning method for neural networks. FiLM layers influence neural network computation via a simple, feature-wise affine transformation based on conditioning information. Inspired by these works, Wang *et al.* [?] proposed a spatial feature transformation (SFT) layer to modulate the features of some intermediate layers in a single network conditioned on semantic segmentation probability maps. Our approach is partially inspired by the above feature normalization methods, which can alter the behavior of deep CNNs to influence the output. In terms of network architecture, we use the Residual-in-Residual Dense Block (RRDB) [?] equipped with SFT layers.

### 2.2.3 Continuous Imagery Effect Transition

Since the restored image’s perceived quality is relatively subjective, and the perception-oriented methods sometimes generate artifacts, users may wish to control the reconstruction result according to the preferences or image characteristics. In recent years, there have been some tunable models that produce intermediate images between the goals of two different objective functions. Specifically, these methods start by training several separate models and then propose different ways of interpolating between them, specifically by directly interpolating the output pixels or network weights [?, ?], or by using specialized adaptor blocks in the networks [?]. They considered trade-off relationships between two objectives, such as perception-distortion balance in SR, noise reduction vs. detail preservation in denoising and style transfer [?, ?, ?, ?]. However, these methods have some limitations: the number of objective functions is two, and they cannot adjust local regions, *i.e.*, the algorithm is equally applied to the entire region of an image. It is also inefficient that they have to train and store multiple separate models. On the other hand, Bahat *et al.* [?] proposed an explorable SR framework that enables local restoration control. However, users have to manually edit the texture in a few steps through a user interface. For easier and more effective quality control, we propose a controllable SR model that can produce various recovery styles for each region with a simple adjustment method. Besides, we can generate intermediate results between two or more different styles at fine control levels.

### 2.2.4 Multi-task Learning

Learning one task at a time is a typical methodology in machine learning because it is hard to simultaneously optimize multiple objectives due to model capacity limitation or conflicting losses. For this reason, such multi-objective problems are commonly scalarized by a linear combination of the losses, with weights defining the trade-off between the loss term [?]. On the other hand, Multi-task Learning (MTL) is an inductive transfer mechanism whose goal is to improve generalization performance by

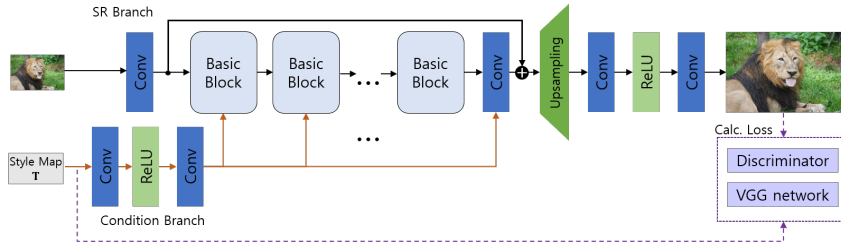


Figure 2.2: The architecture of our proposed flexible SR network. We use the RRDB equipped with SFT as a basic block (Figure ??(c)). The condition branch takes a style map for reconstruction style as input. This map is used to control the recovery styles of edges and textures for each region through SFT layers.

leveraging useful domain-specific information contained in multiple related tasks [?]. Specifically, since the MTL networks use shared layers trained in parallel on all the tasks, what is learned for each task can help others to learn better when tasks are closely related [?,?]. Recently, Dosovitskiy *et al.* [?] proposed loss-conditional training of deep networks for MTL that can improve model efficiency by exploiting the redundancy of multiple related models. They demonstrate style-transfer trained in this way and utilize feature-wise linear modulation [?] that affects the whole image style.

## 2.3 Proposed Method

### 2.3.1 Targeted Perceptual Loss

In general, the choice of feature space significantly influences perceptual reconstruction performance and the styles. For example, Figure ?? shows the effect of choosing different feature spaces in computing the perceptual loss. In this paper, four different layers, ReLU 2-2, ReLU 3-4, ReLU 4-4, and ReLU 5-4 of the VGG-19 network [?] are considered, denoted as VGG22, VGG34, VGG44, and VGG54, respectively. As shown in Figure ??, while the low-level feature space VGG22 seems more suitable for reconstructing simple edges with less distortion and over-sharpening, the mid- and

high-level feature spaces of VGG44 are more appropriate for recovering complex textures. Therefore, it is difficult to determine a single feature space that works best for the entire image.

In our work, we use more than two feature spaces at the same time to train a flexible SR (FxSR) model capable of generating various reconstruction styles. We define two kinds of FxSR models, namely FxSR-PD (perception-distortion) and FxSR-DS (diversity). The FxSR-PD is the main model in our work, which controls the output style between the distortion-oriented and perception-oriented by combining the reconstruction loss (for distortion) and VGG22 feature loss (for perception), along with the adversarial loss. The FxSR-DS uses the same architecture as the FxSR-PD but is trained with different losses, including all the VGG features stated above. Hence, the aim of FxSR-DS is to produce diverse styles of outputs related to different VGG features rather than to control between distortion and perception. Unlike previous works where there is no control data, we adjust the network by applying different objective functions for each local region through a style control map<sup>1</sup>. As a result, we can explore various HR solutions that are generated using multiple objective functions and thus reconstruct an image with the desired style or an image closer to the original HR.

### 2.3.2 Proposed SR with Flexible Style

Given a single LR image  $I^{LR}$ , SISR is to estimate an HR image  $\hat{I}^{HR}$ , which is as similar as possible to its corresponding HR counterpart  $I^{HR}$ . Most of the current CNN-based methods use feed-forward networks to directly learn a mapping function  $G_\theta$  parameterized by  $\theta$  as

$$\hat{I}^{HR} = G_\theta (I^{LR}). \quad (2.1)$$

To optimize  $G_\theta$  on the training samples, we design a specific objective function  $\mathcal{O}$  as

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{\mathcal{Z} \sim \mathcal{P}_{\mathcal{Z}}} \left[ \mathcal{O} \left( \hat{I}^{HR}, I^{HR} \right) \right] \quad (2.2)$$

---

<sup>1</sup>In the rest of the paper, we will refer the style control map as just *style map* or a map T.

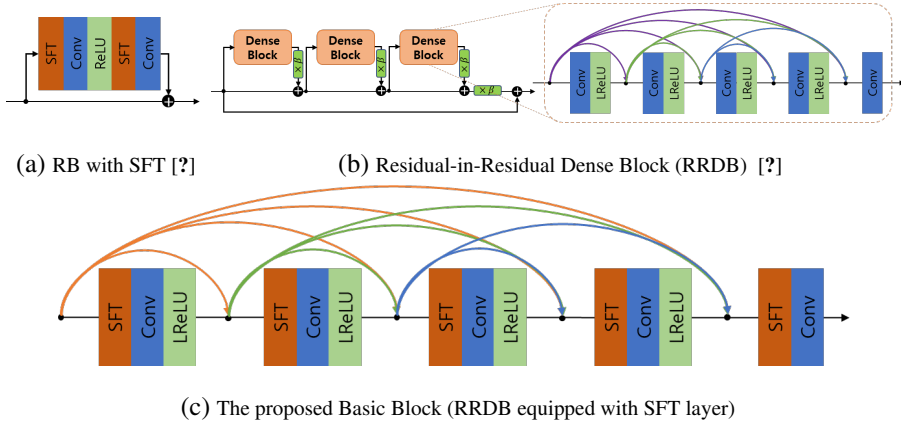


Figure 2.3: RRDB with SFT for basic blocks

where  $\mathcal{Z} = (I^{LR}, I^{HR})$  is sampled from given a training distribution of pairs  $P_{\mathcal{Z}}$ . Many recent studies [?, ?] use perceptual loss and adversarial loss for designing  $\mathcal{O}$  to recover realistic textures. Although these losses greatly improve the perceptual quality, the generated textures tend to be monotonous and unnatural [?, ?]. To further improve the restoration performance, Wang *et al.* [?] used semantic segmentation probability maps as the categorical prior  $\Psi$  and reformulated (??) as

$$\hat{I}_{\Psi}^{HR} = G_{\theta} (I^{LR} | \Psi). \quad (2.3)$$

However, the perceptual loss was applied to the entire region of images, like in previous works. Specifically, the same level of features was used both on simple edges and complex textures, which has a limitation in restoring images composed of various types of objects. In addition, once model training is completed, there is no way to adjust the SR results without retraining. Hence, instead, we propose a novel method to apply different objectives to each region for reconstructing desired images or images closer to the original. Specifically, the proposed flexible SR model is optimized with a conditional objective, which is a weighted sum of several perceptual losses corresponding to different feature levels, where each weight changes depending on the style map.

Formally, our objective is described as:

$$\hat{I}_{\mathbf{T}}^{HR} = G_{\theta} (I^{LR} | \mathbf{T}), \quad (2.4)$$

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{t \sim \mathcal{P}_t} \mathbb{E}_{\mathcal{Z} \sim \mathcal{P}_{\mathcal{Z}}} \left[ \mathcal{O} \left( \hat{I}_{\mathbf{T}}^{HR}, I^{HR} | \mathbf{T} \right) \right] \quad (2.5)$$

where  $\mathbf{T}$  is a map delivering spatially varying style control. That is, the map  $\mathbf{T}$  is an LR-sized matrix, which is fed to the condition network to change the SR styles. Since the purpose of training is to let the network learn various styles corresponding to given control parameters, we feed various  $\mathbf{T}$  randomly to the network during the training. Specifically, we feed a flat map  $\mathbf{T} = t \times \mathbf{1}$  during the training, where  $\mathbf{1}$  is the matrix with all the elements 1, and  $t$  is a variable related to the feature combinations, which will be detailed in the following subsection. For training with various feature combinations, we change  $t$  randomly at each epoch. At the inference, if we feed a flat map as defined above, the network will deliver an SR style globally corresponding to the  $t$ . If we wish to control the styles locally, we feed a spatially varying map, which will be demonstrated in the experiment.

### 2.3.3 Proposed Network Architecture

An overview of the architecture is shown in Figure ???. The generator network  $G_{\theta}$  consists of two streams, an SR branch and a condition branch. The SR branch is built with basic blocks consisting of RRDB equipped with the SFT layers [?], which take the shared conditions as input and modulate feature maps by applying the affine transformation. This structure is shown in Figure ??(c), where the residual block with SFT [?] and RRDB [?] are also shown in Figures ??(a) and (b) for comparison. The SFT layer learns a mapping function that outputs a modulation parameter based on a style condition  $\mathbf{T}$ . This modulation layer allows the SR branch to optimize the changing objective during the training and also to generate SR results with spatially different styles according to the style map. The condition branch is used to produce shared intermediate

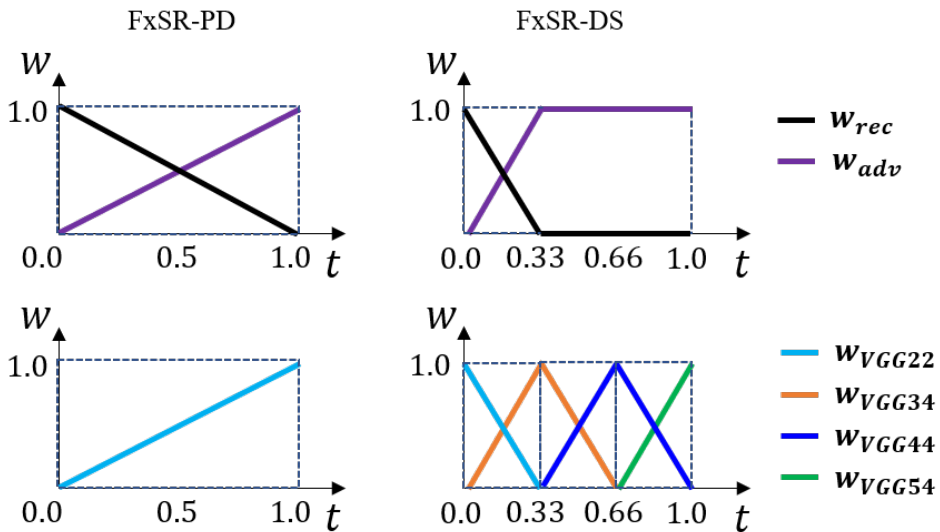


Figure 2.4: The left column shows the weight functions for FxSR-PD.  $t = 0$  corresponds to distortion-oriented SR (only MSE loss) and  $t = 1$  perception-oriented (with adversarial and perceptual loss from VGG22). The right column shows the weight functions for FxSR-DS, where more perceptual losses are used to expand the HR styles.

style conditions that can be broadcasted to all the SFT layers for efficiency. As in the study of [?], all the convolution layers in the condition branch are restricted to use  $1 \times 1$  kernels to avoid the interference of different regions. For discriminator network, we use VGG network [?] that contains ten convolution layers gradually decreasing the spatial dimensions.

### 2.3.4 Proposed Loss Function

We combine multiple losses to train our SR model. The conditional objective consists of three terms, namely pixel-wise reconstruction loss, adversarial loss, and proposed conditional perceptual loss:

$$\mathcal{O}_{\mathbf{T}} = \lambda_{rec}(\mathbf{T}) \cdot \mathcal{L}_{rec} + \lambda_{adv}(\mathbf{T}) \cdot \mathcal{L}_{adv} + \lambda_{per} \cdot \mathcal{L}_{per}(\mathbf{T}) \quad (2.6)$$

where

$$\lambda_{rec}(\mathbf{T}) = \lambda_{rec_o} + (\eta \cdot w_{rec}(\mathbf{T})), \quad (2.7)$$

$$\lambda_{adv}(\mathbf{T}) = \lambda_{adv_o} \cdot w_{adv}(\mathbf{T}). \quad (2.8)$$

The notations will be explained one by one below. First, the reconstruction loss is calculated as:

$$\mathcal{L}_{rec} = \mathbb{E} \left[ \|\hat{I}^{HR} - I^{HR}\|_1 \right]. \quad (2.9)$$

We use the adversarial loss using Relativistic average Discriminator RaD [?] that performs better for learning sharper edges and more detailed textures compared to standard GAN [?]. While the standard version estimates the probability that one input image  $I$  is real and natural, the RaD predicts the probability that a real image  $I^{HR}$  is relatively more realistic than a fake one  $\hat{I}^{HR}$ . In addition, for adversarial training, RaD benefits from the gradients from both  $\hat{I}^{HR}$  and  $I^{HR}$ , while only  $\hat{I}^{HR}$  takes effect in the standard version. Specifically, the adversarial and the discriminator losses are:

$$\mathcal{L}_{adv} = -\mathbb{E}_{\hat{I}^{HR}} \left[ \log \left( \tilde{D} \left( \hat{I}^{HR} \right) \right) \right] - \mathbb{E}_{I^{HR}} \left[ \log \left( 1 - \tilde{D} \left( I^{HR} \right) \right) \right] \quad (2.10)$$

$$\mathcal{L}_{dis} = -\mathbb{E}_{I^{HR}} \left[ \log \left( \tilde{D} \left( I^{HR} \right) \right) \right] - \mathbb{E}_{\hat{I}^{HR}} \left[ \log \left( 1 - \tilde{D} \left( \hat{I}^{HR} \right) \right) \right] \quad (2.11)$$

where

$$\tilde{D} \left( I^{HR} \right) = \text{sigmoid} \left( C \left( I^{HR} \right) - \mathbb{E}_{\hat{I}^{HR}} \left[ C \left( \hat{I}^{HR} \right) \right] \right) \quad (2.12)$$

$$\tilde{D} \left( \hat{I}^{HR} \right) = \text{sigmoid} \left( C \left( \hat{I}^{HR} \right) - \mathbb{E}_{I^{HR}} \left[ C \left( I^{HR} \right) \right] \right) \quad (2.13)$$

where  $C(\cdot)$  represents the output logit of discriminator.

The conditional perceptual loss is a weighted sum of multiple perceptual losses in different levels of feature spaces:

$$\mathcal{L}_{per}(\mathbf{T}) = \sum_l w_l(\mathbf{T}) \cdot \mathcal{L}_l, \quad (2.14)$$



where  $\mathcal{L}_l$  denotes the distance in each feature space,  $l \in \{VGG12, VGG22, \dots, VGG54\}$ , and the weights  $w_l$  changes according to  $\mathbf{T}$ . Precisely, the distance  $\mathcal{L}_l$  is defined as

$$\mathcal{L}_l = \mathbb{E} \left[ \|\phi_l(\hat{I}^{HR}) - \phi_l(I^{HR})\|_2 \right] \quad (2.15)$$

where  $\phi_l$  denotes feature maps in the feature space  $l$ . The weights  $w_{rec}$ ,  $w_{adv}$ , and  $w_l$  are functions of  $t$  as described in Figure ??, where  $t$  is a random variable having uniform distribution in  $[0, 1]$  during the training.

### 2.3.5 Implementation details

This subsection explains how we design the combination of feature losses depending on the change of  $t$ . The left column of Figure ?? shows the weight function for FxSR-PD (using only VGG22 for perceptual loss), and the right for FxSR-DS (using more feature spaces for diversity). When  $t=0$ , the figure shows that FxSR-PD corresponds to distortion-oriented SR (perceptual and adversarial losses are zero). When the value of  $t$  approaches 1, then it becomes perception-oriented (weight for the reconstruction loss becomes zero, while adversarial and perceptual losses grow to 1). In the case of the right column, various feature distances are involved in the perceptual loss, and hence FxSR-DS can deliver diverse styles. Specifically, note that  $t = 1$  corresponds to a perception-oriented SR with VGG54 as the feature space. Also, even when  $t$  approaches 0, the FxSR-DS still produces perception-oriented SR results of different styles corresponding to VGG22, unlike the FxSR-PD that is distortion-oriented at  $t = 0$ .

Regarding the style control, as stated previously, we use a uniform map  $\mathbf{T} = t \times \mathbf{1}$  at the training phase. That is, a flat map is fed to the condition branch, with its intensity  $t$  randomly changing during the training. Since the SR network is a fully convolutional neural network, it inherits the local connectivity property that the local image and the map region determine the output pixel. Hence, SR models trained with uniform maps can handle spatially varying cases.

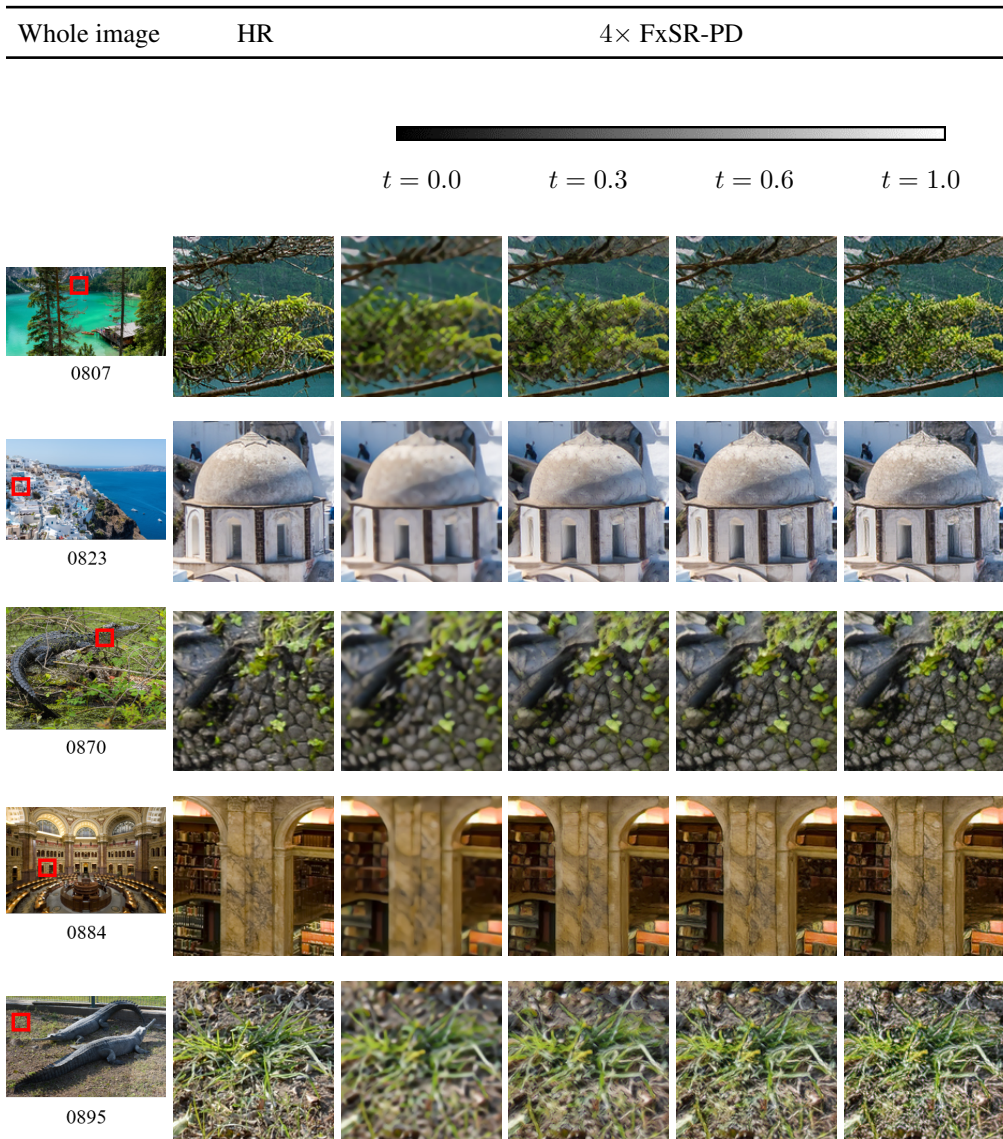


Figure 2.5: Changes in the result of FxSR-PD 4× SR according to  $t$  on DIV2K validation set [?].

## 2.4 Experiments

In the experiment, we compare our FxSR-PD and FxSR-DS with several state-of-the-art SR methods on benchmark datasets. We start the section with a description of the

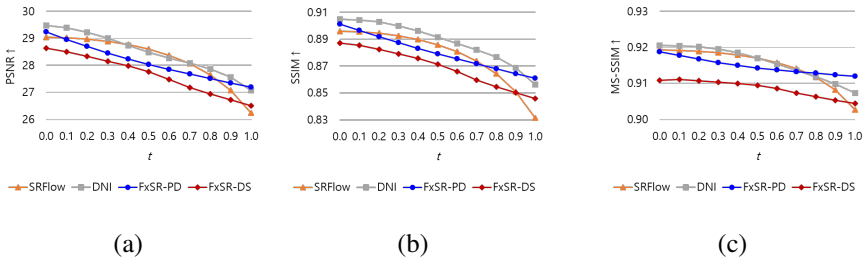


Figure 2.6:  $4\times$  SR performance comparison of state-of-the-art and proposed methods evaluated by the metric (a) PSNR, (b) SSIM [?], and (c) MS-SSIM [?] for DIV2K according to condition parameters.

datasets and evaluation methods. Next, we present the comparison results. We also provide examples of local style control and validate the effectiveness of our approach for compressed images. Finally, we report complexity analysis for the proposed methods.

## 2.4.1 Materials and Methods

### Datasets

For the experiments, we train the FxSR with DIV2K [?] dataset, which contains 800 training images, 100 validation images, and 100 test images. We use BSDS100, General100, and DIV2K 100 validation images as our test datasets. We also use JPEG-compressed images for training and testing FxSR models to show that our proposed method is still effective on the real-world compressed LR images. The scaling factors of  $4\times$  and  $8\times$  are tried for experiments.

### Evaluation Method

To evaluate the perceptual distance to the Ground Truth, we report LPIPS [?] as default [?], and additionally use DISTS [?] as structure and texture similarity in some cases. PSNR and SSIM [?] are reported as fidelity-oriented metrics. Furthermore, we report the no-reference metric NIQE [?]. Since the consistency with the LR image is also

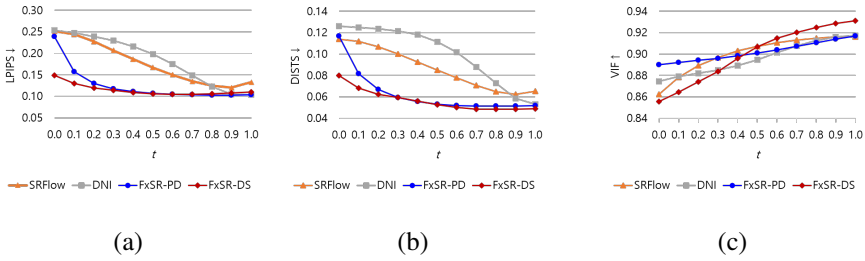


Figure 2.7:  $4\times$  SR performance comparison of state-of-the-art and proposed methods evaluated by the metric (a) LPIPS [?], (b) DISTS [?], and (c) VIF [?] for DIV2K according to condition parameters.

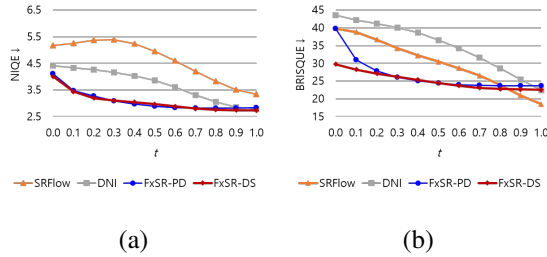


Figure 2.8:  $4\times$  SR performance comparison of state-of-the-art and proposed methods evaluated by the (a) NIQE [?] and (b) BRISQUE [?] for DIV2K according to condition parameters.

an important factor, we report the LR-PSNR, computed as the PSNR between the downsampled SR image and the original LR. To measure the meaningful diversity of SR methods that can actively sample from the space of plausible super-resolutions, we also report the SR-Diversity score, which is used for the evaluation protocol on the Super-Resolution Space Challenge learning track in the NTIRE Challenge 2021 [?,?]. Specifically, we sample 11 images and densely calculate LPIPS [?] metric between the samples and the ground truth. To obtain the local best score, we pixel-wisely select the best score out of the 11 samples and take the full image’s average. The global best score is calculated by averaging the whole image’s score and selecting the best. Then,

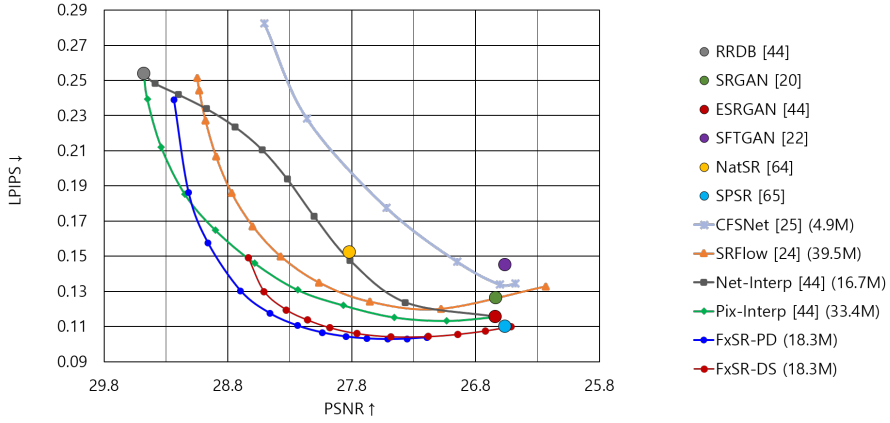


Figure 2.9: Performance comparison of the state-of-the-arts and proposed method (FxSR-PD: Perception-Distortion Flexible SR) for DIV2K  $4\times$  SR.

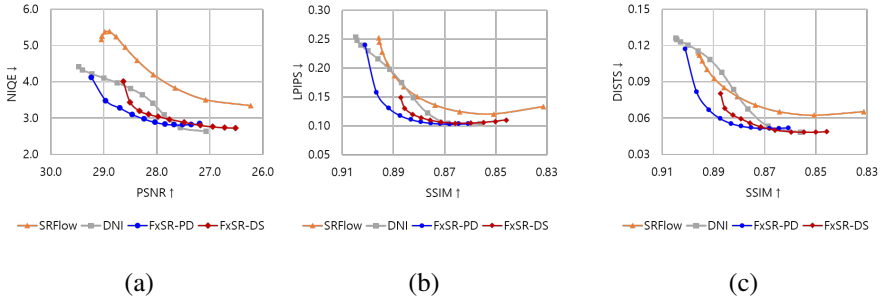


Figure 2.10:  $4\times$  SR P-D performance comparison of state-of-the-art and proposed methods evaluated by the metric (a) NIQE [?] Vs. PSNR, (b) SSIM [?] Vs. LPIPS [?], and (c) SSIM [?] Vs. DISTS [?] for DIV2K according to condition parameters.

the diversity score is calculated as follows:

$$score = (globalbest - localbest)/(globalbest) \times 100. \quad (2.16)$$

## Training Method

For the scaling factor  $4\times$ , sub-images are cropped with the sizes of  $320 \times 320$  with a stride of 160 and  $80 \times 80$  with 40, for the HR and LR training images, respectively.

Table 2.1: Comparison with state-of-the-art SR methods on benchmarks. In  $4\times$  and  $8\times$ , the 1st and the 2nd best performances are highlighted in **bold** and underline, respectively.

Dataset	Metric	$4\times$										$8\times$						
		RRDB	SRGAN	ESR-GAN	SFT-GAN	NatSR	SPSR	SRFlow	SRFlow	FxSR-PD	FxSR- $t=0.8$	RRDB	ESR-GAN	SRFlow	SRFlow	FxSR-PD	FxSR- $t=0.0$	
BSD100	PSNR $\uparrow$	<b>26.53</b>	24.13	23.95	24.09	25.13	24.16	24.66	26.23	24.66	<u>26.38</u>	24.77	23.56	20.23	23.37	21.66	<b>23.60</b>	21.93
	SSIM $\uparrow$	<b>0.7438</b>	0.6454	0.6463	0.6460	0.6780	0.6531	0.6580	0.7293	0.6580	<u>0.7380</u>	0.6817	<b>0.5700</b>	0.4350	0.5428	0.4632	<u>0.5728</u>	0.5039
	LrPSNR $\uparrow$	<u>51.52</u>	39.32	41.35	40.92	42.26	40.99	50.81	49.86	50.81	<b>52.48</b>	49.24	45.82	24.81	<b>52.39</b>	<u>51.09</u>	47.12	42.41
	LPIPS $\downarrow$	0.3575	0.1777	0.1615	0.1710	0.2115	<u>0.1613</u>	0.3635	0.1833	0.3635	0.3433	<b>0.1572</b>	0.5571	0.3582	0.5303	<u>0.3238</u>	0.5079	<b>0.3129</b>
	DISTS $\downarrow$	0.2005	0.1288	<b>0.1160</b>	0.1224	0.1436	0.1165	0.1943	0.1372	0.1943	0.1921	<b>0.1160</b>	0.2956	0.2096	0.3183	<u>0.2068</u>	0.2753	<b>0.1972</b>
	NIQE $\downarrow$	5.35	<b>3.18</b>	3.53	3.23	3.67	3.23	6.83	6.83	5.10	3.30	6.23	<b>3.15</b>	12.82	3.68	5.49	4.58	
General100	PSNR $\uparrow$	<b>30.30</b>	27.54	27.53	27.04	28.61	27.65	27.83	29.72	27.83	29.94	28.44	<u>25.38</u>	21.51	25.09	23.45	<b>25.42</b>	24.00
	SSIM $\uparrow$	<b>0.8696</b>	0.7998	0.7984	0.7861	0.8259	0.7995	0.8574	0.874	0.7951	<u>0.8629</u>	0.8229	<u>0.7081</u>	0.5674	0.6806	0.6063	<b>0.7097</b>	0.6534
	LrPSNR $\uparrow$	<b>53.96</b>	41.44	41.93	40.05	45.06	42.31	50.65	49.59	50.65	<u>52.22</u>	49.82	44.78	25.19	<b>48.95</b>	<u>47.59</u>	44.28	41.36
	LPIPS $\downarrow$	0.1665	0.0962	0.0881	0.1084	0.1118	<u>0.0865</u>	0.1731	0.0962	0.1731	0.1519	<b>0.0784</b>	0.3403	0.2494	0.3194	<u>0.2341</u>	0.2924	<b>0.2058</b>
	DISTS $\downarrow$	0.1321	0.0955	<u>0.0845</u>	0.1166	0.1099	0.0857	0.1276	0.1022	0.1276	0.1205	<b>0.0831</b>	0.2362	<u>0.1852</u>	0.2488	0.1899	0.2134	<b>0.1716</b>
	NIQE $\downarrow$	6.56	<b>4.35</b>	4.65	4.38	4.71	4.37	7.02	7.02	6.05	4.54	7.18	<b>4.40</b>	11.92	4.89	6.09	5.46	
DIV2K	PSNR $\uparrow$	<b>29.48</b>	26.63	26.64	26.56	27.82	26.71	27.08	29.05	27.08	29.24	27.51	<u>25.50</u>	21.37	25.09	23.04	<b>25.60</b>	23.56
	SSIM $\uparrow$	<b>0.8444</b>	0.7625	0.7640	0.7578	0.7931	0.7614	0.8290	0.8290	0.7558	<u>0.8383</u>	0.7890	<u>0.6951</u>	0.5533	0.6589	0.5728	<b>0.6989</b>	0.6241
	LrPSNR $\uparrow$	<b>53.72</b>	40.87	42.61	40.40	44.64	42.57	51.02	49.96	49.96	<u>53.30</u>	50.54	46.05	25.21	<b>51.28</b>	<u>50.26</u>	46.96	42.66
	LPIPS $\downarrow$	0.2537	0.1263	0.1154	0.1449	0.1523	<u>0.1099</u>	0.2513	0.1201	0.2390	0.2390	<b>0.1028</b>	0.4245	0.2841	0.4033	<u>0.2719</u>	0.3857	<b>0.2403</b>
	DISTS $\downarrow$	0.1261	0.0613	0.0530	0.0858	0.0766	<b>0.0493</b>	0.1139	0.0622	0.1169	0.1169	<u>0.0513</u>	0.2203	<u>0.1293</u>	0.2342	0.1386	0.1953	<b>0.1190</b>
	NIQE $\downarrow$	4.42	<b>2.57</b>	2.79	2.92	2.91	2.74	5.16	5.16	4.11	2.81	5.15	<b>2.53</b>	7.15	3.54	4.41	3.61	
Param.		16.7M	1.5M	16.7M	53.7M	4.8M	24.8M	39.5M	39.5M	18.3M	18.3M	16.7M	16.7M	16.7M	50.8M	50.8M	18.3M	18.3M

4× SR comparison




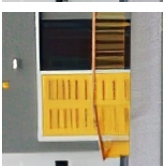
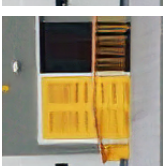
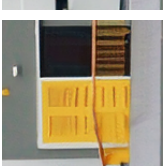
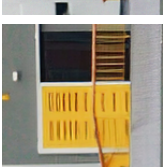
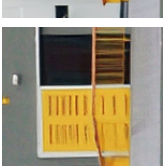


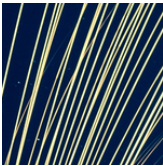
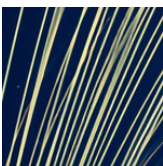
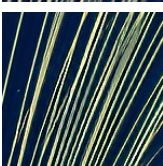
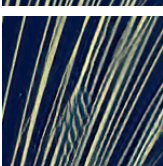
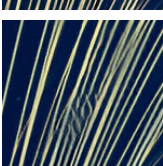
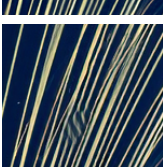
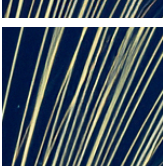
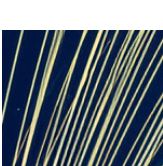


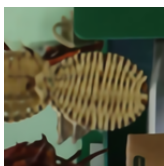
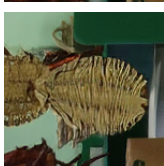
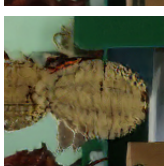
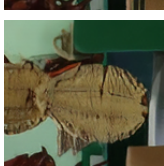
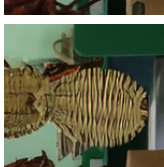
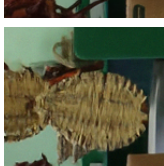
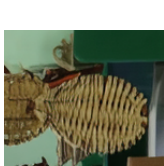


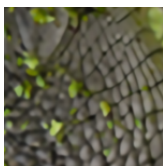
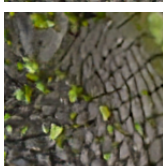
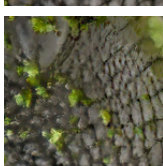
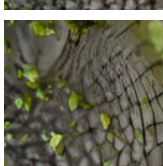
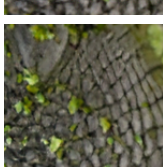
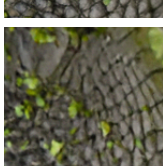
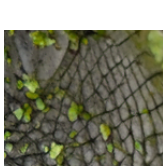



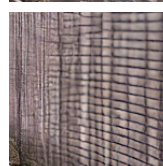
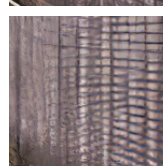
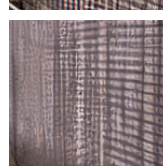
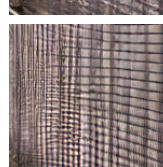

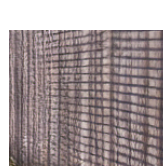
Whole image	HR	RRDB [?]	ESRGAN [?]	SFTGAN [?]	NatSR [?]	SPSR [?]	SRRFlow $t=0.9$	FxSR $t=0.8$
 0836								
 0828								
 0837								
 0870								
 0876								

Figure 2.11: Visual comparison with state-of-the-art perception-driven SR methods on DIV2K validation set [?]. The proposed method produces competitive results compared to other modern techniques and can also generate reconstructed images of various styles of LR images.

### 8× SR comparison

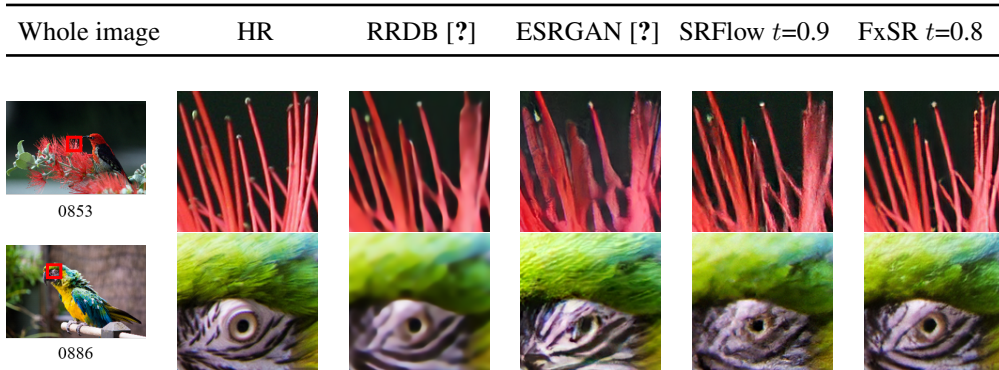


Figure 2.12: Visual comparison for 8× SR results on DIV2K validation set [?].

For the scaling factor 8×, the LR sub-images are cropped to the size of  $40 \times 40$  with a 20 strides. Then, the batch image pairs for each iteration of training are randomly cropped from these sub-images. The HR batch size is  $128 \times 128$  and the LR batch sizes are  $32 \times 32$  and  $16 \times 16$  for scaling factors of 4× and 8×, respectively.

For the optimization, we use initial learning rate of  $10^{-4}$ . The learning rate is halved after 5K, 10K, 20K, and 30K iterations. Adam [?] with  $\beta_1 = 0.9$  and  $\beta_2 = 0.99$  is used for both generator and discriminator training. We use pre-trained RRDB [?] and ESRGAN [?] models to optimize the proposed FxSR models. While fine-tuning FxSR-PD and FxSR-DS,  $\lambda_{rec.o}$ ,  $\lambda_{adv.o}$  and  $\lambda_{per}$  are set to be  $1 \times 10^{-2}$ ,  $5 \times 10^{-3}$  and 1.0 respectively, but  $\eta$  is set differently to  $1 \times 10$  and 1.0.

#### 2.4.2 Evaluation of Flexible SR for Perception-Distortion (FxSR-PD)

By adjusting a single parameter  $t$ , the FxSR-PD model can generate various SR results for the trade-offs between distortion and perception objective at the inference phase, as shown in Figure ?? . It shows that  $t = 0$  generates blurry outputs as the FxSR objective is distortion-oriented, and  $t = 1$  generates sharp textures as the FxSR becomes perception-oriented. Also, the  $t$  between 0 and 1 generates different trade-offs, with less or more distortions, and more or less blurriness.



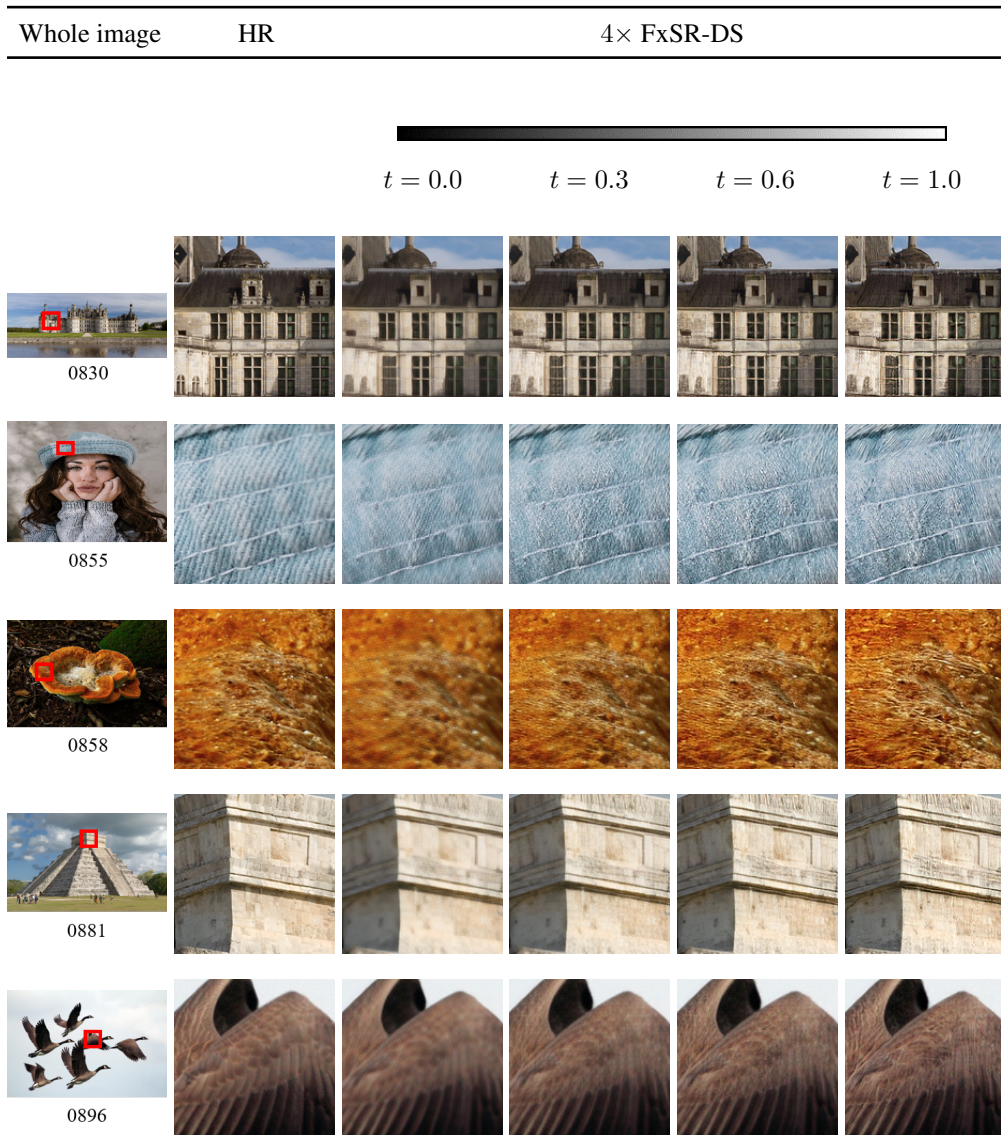


Figure 2.13: Changes in the result of FxSR-DS 4× SR according to  $t$  on DIV2K validation set [?].

### Quantitative Comparison

We compare our method quantitatively with distortion-oriented methods such as RRDB [?], and perception-oriented methods such as SRGAN [?], ESRGAN [?], SFTGAN

[?], NatSR [?], SPSR [?] and SRFlow [?]. For the  $4\times$  SR, we use pre-trained models provided by the authors, while for the non-provided  $8\times$  SR, we used the author’s code to train the RRDB [?] and ESRGAN [?] models. The results are presented in Figures from ?? to ?? and Table ?. Figures from ?? to ?? show the performance comparison of  $4\times$  SR results according to  $t$ , evaluated by the distortion-oriented (PSNR, SSIM [?], MS-SSIM [?]), perception-oriented (LPIPS [?], DISTS [?], VIF [?]), and non-reference perception-oriented metrics (NIQE [?], BRISQUE [?]), respectively. In Figure ??, we can see that the scores of the distortion-oriented metrics improve as  $t$  approaches 0, whereas in Figures ?? and ??, the scores of the perception-oriented metrics improve as  $t$  approaches 1.

Since there is a trade-off between the distortion-oriented metrics and the perception-oriented metrics, it is necessary to evaluate the performance of the SR models in a perception-distortion 2D plane [?], as shown in Figure ?. The vertical axis denotes perceptual loss LPIPS [?], and the horizontal axis the PSNR (distortion-oriented measure). Hence, the lower left part is the desired place where both MSE and perceptual loss are low [?], and we can see that our method is comparable to others in this respect. Note that the RRDB [?] and ESRGAN [?] are the results of using distortion-oriented and perception-oriented loss, respectively. Others drawn in solid lines are adjustable methods. Pixel interpolation (Pix-Interp) and network weight interpolation (Net-Interp) methods utilize two differently trained models, *i.e.*, the RRDB and ESRGAN stated above. The number of parameters for each method is also provided for complexity comparison. More details about complexity analysis will be provided in Section IV.F.

Since various metrics examined in Figures ??- ?? have different characteristics and performance, we present additional performance comparisons for the perception-distortion plane with these metrics in Figure ?. These comparisons show trends similar to those in Figure ?. Table ? shows the evaluation of FxSR-PD and other SR methods for the specific  $t$  values. The proposed FxSR-PD obtains the best PSNR

and SSIM at  $t = 0$  among perception-oriented methods and the best LPIPS values at  $t = 0.8$  for all datasets.

Table 2.2: Comparison with state-of-the-art SR methods on DIV2K in terms of low resolution consistency, photo-realism and meaningful diversity. The numbers in the parentheses are the relative performances, i.e., the best value is set to 1, and the others are divided by the best value.

	SR Model	LR-PSNR $\uparrow$	Mean LPIPS $\downarrow$	G-best LPIPS $\downarrow$	L-best LPIPS $\downarrow$	Div. score $\uparrow$
4 $\times$	SRFlow [?]	50.55 (0.99)	0.1765 (1.54)	0.1153 (1.14)	0.0905 (1.03)	<b>23.12</b> <b>(1.00)</b>
	DNI [?]	44.37 (0.87)	0.1968 (1.72)	0.1114 (1.10)	0.1003 (1.14)	10.01 (0.43)
	FxSR-PD	<b>51.16</b> <b>(1.00)</b>	0.1253 (1.10)	<b>0.1010</b> <b>(1.00)</b>	0.0926 (1.05)	8.98 (0.39)
	FxSR-DS	44.49 (0.87)	<b>0.1144</b> <b>(1.00)</b>	0.1018 (1.01)	<b>0.0880</b> (1.00)	13.66 (0.59)
8 $\times$	SRFlow [?]	<b>50.78</b> (1.00)	0.3261 (1.32)	0.2613 (1.19)	0.2066 (1.08)	<b>21.88</b> <b>(1.00)</b>
	FxSR-PD	44.76 (0.88)	<b>0.2477</b> <b>(1.00)</b>	<b>0.2192</b> <b>(1.00)</b>	0.1996 (1.04)	9.11 (0.42)
	FxSR-DS	37.77 (0.74)	<b>0.2477</b> <b>(1.00)</b>	0.2206 (1.01)	<b>0.1912</b> <b>(1.00)</b>	13.39 (0.61)

## Qualitative Comparison

Visual comparison between our proposed FxSR-PD and other state-of-the-art methods for 4 $\times$  and 8 $\times$  are shown in Figures ?? and ??, respectively. We can see that our FxSR-PD provides stronger edges and fine details than the distortion-oriented method RRDB [?], and other perception-oriented ones. Also, there are fewer artifacts in our method compared to others.

### 2.4.3 Flexible SR for Diverse Styles (F<sub>x</sub>SR-DS)

#### Diverse Style HR Generation

Unlike the F<sub>x</sub>SR-PD that attempts flexible trade-offs between perception and distortion, the F<sub>x</sub>SR-DS aims to generate various styles of HR textures with perceptually high scores for all  $t$  values. As shown in Figures from ?? to ??, the F<sub>x</sub>SR-DS scores better overall with a relatively narrow dynamic range regarding the perception-oriented metrics other than VIF [?]. On the other hand, it scores relatively lower for distortion-oriented metrics as in Figure ?. The loss terms and their weights for the conditional objective of the F<sub>x</sub>SR-DS model are described in Figure ?. Different from F<sub>x</sub>SR-PD with one perceptual loss term, four perceptual loss terms at different feature levels are used. In Figure ??, we can see that the SR results for different  $t$  values have different types of styles that are clearly distinct from each other. While Figure ?? shows the trade-off results between perception and distortion, Figure ?? visualizes our method’s scalability to generate various styles of textures by employing more feature spaces into the loss.

#### Quantitative Comparison

Table ?? compares with DNI [?] and SRFlow [?] in terms of LRPSNR (low-resolution PSNR), LPIPS and Diversity metrics which are evaluation protocol on the Ntire 2021 Challenge [?, ?] stated previously. Table ?? is the evaluation of SR results for a specific  $t$  value, while Table ?? is the average of all of the SR results for 11 different  $t$  values, from 0 to 1, with the step size of 0.1. Specifically, in Table ??, the F<sub>x</sub>SR-DS generally scores the best mean LPIPS and Local best (L-best) LPIPS, while the F<sub>x</sub>SR-PD achieves the best Global best (G-best) LPIPS score. This proves that the perceptually distinct diverse SR results generated by F<sub>x</sub>SR-DS in Figure ?? are of high quality in terms of perception-oriented metrics. Since Local Best LPIPS is the maximum performance of the SR model in terms of perceptual measurement, the

proposed FxSR-DS shows an improvement of about 2.7% compared to the SRFlow. Figure ?? also demonstrates that while the FxDR-PD scores better G-best LPIPS compared to FxDR-DS, the FxDR-DS scores rather superior L-best LPIPS than FxSR-PD. Meanwhile, the SRFlow [?] produces the highest diversity, which learns the sample distribution during training while the proposed models are trained to optimize objectives in the training distribution of objective. However, it is also important to note that the diversity scores are normalized by the G-best as Eqn. ?. This means that the higher the G-best LPIPS, that is, the lower the absolute perceptual quality level, the higher the diversity score.

#### 2.4.4 Per-pixel Style Control

In this section, we demonstrate some examples of applying local style control. First, Figure ?? is an example where the LR image has both text and texture areas. In the conventional methods for the SR of Figure ??(a), multiple SR models are trained with one objective each. Then a model is selected, and the entire image is optimized with the model’s objective. If the SR model 0 is selected, which is RRDB [?] representing the distortion-oriented model, the textures of the clothes are blurred while the text edges are restored without artifacts. Conversely, suppose we select the SR model  $N - 1$ , which is ESRGAN [?] representing the perception-oriented model. In that case, some characters in the text area are broken while the textures of the clothes are naturally restored. On the other hand, the proposed FxSR-PD in Figure ??(b) can restore both the textures of clothes and characters at the same time by applying different objectives to each area through the locally-manipulated style map.

As the second example, let us consider the structural edges of the building and textures of the tree area in Figure ?. In a typical approach of using multiple SR models in Figure ??(a), when the SR model 0 (RRDB) is selected, the structural edges of the building are restored without artifacts, but the tree textures are blurred. Conversely, if the SR model  $N-1$  (ESRGAN) is chosen, the overshoot side-effect occurs around the

edges. As shown in Figure ??(b), similar to the previous example, when a properly adjusted local style map is fed along with the input image, the proposed model FxSR-DS can restore both the tree textures and building edges naturally.

The next is an example of enhancing the perspective feeling when depth information is available, as shown in Figure ?. Precisely, input image and depth map pairs used in this example are from the Make3D data set [?, ?]. When the distance map is used as  $\mathbf{T}$  in our FxSR, the foreground region is super-resolved in a perception-oriented way (with emphasized texture), and the background region is distortion-oriented (somewhat blurry). Depth information obtained by some equipment such as Kinect [?] and Time-of-Flight (ToF) camera [?, ?], or depth estimation algorithms [?] can be used. It is also possible for users to directly generate a depth map from an input image using image editing S/W, as shown in Figure ?. This makes the foreground clearer with sharp details and avoids the unnaturalness of the background becoming as sharp as the foreground. In addition, the camera noise in the background can be reduced. As seen in the examples so far, the proposed method can be used for most cases in various fields that require different processing for each area for a specific purpose.

#### 2.4.5 Compressed LR Image Restoration

Since real-world SR is challenging due to unknown degradation and various noise [?, ?, ?, ?, ?, ?], we also validate the effectiveness of our method for compressed inputs in Figure ?. Unlike previous experiments, FxSR and SRGAN [?] are re-trained using LR images compressed with JPEG quality factor 90, called FxSR-CA (compression artifacts) and SRGAN-CA. We can see that while compression artifacts are amplified in the results of SRResNet [?] and SRGAN [?] trained with clean images, the proposed FxSR-CA, generates different style and details according to the change of  $t$ . To test the effectiveness of the proposed method for the case of real-world compressed

images, two videos <sup>2</sup> which are filmed, edited and copyrighted by Milosh Kitchovitch are used by courtesy of him. Details of the video are provided in the Table ??.

Table 2.3: The details of video we used for SR performance comparison.

Title	Resolution	Bitrate/Codec
Amazing Place 2018	640×360	319kbps/VP9
Amazing Place 2019	640×360	301kbps/VP9

Table 2.4: Comparisons of the running time, the computational costs and the size of SR models for super-resolution  $4\times$ , when the size of LR input images is  $128 \times 128$ .

	Run Time (msec)	Mult-Add # (G)	Param Size (MB)	Forward Pass (MB)
SRGAN [?]	0.014	1.51	41.63	585.11
ESRGAN [?]	0.138	16.69	293.97	2061.50
FxSR	0.501	18.30	320.20	8432.78

## 2.4.6 Complexity Analysis

We compare the running time, computation costs, and storage size of our methods with other SR methods in Table ?. We measure the complexity for the SR  $4\times$  processing of one  $128 \times 128$  LR input image on the environment of NVIDIA RTX3090 GPU. According to Table ?, ESRGAN with high-complexity RRDB architecture in Figure ??(b) requires about 10 times the number of Mult-Add and Run-time than SRGAN. Compared to ESRGAN, FxSR with the proposed RRDBs with SFT in Figure ??(c) has

<sup>2</sup>URLs of Amazing Place 2018 and Amazing Place 2019: <https://www.youtube.com/watch?v=37IqCYVUhcs>, <https://www.youtube.com/watch?v=g5hA2qo2EFc>

almost the same number of Mult-Adds and parameter size, but the Forward Pass Size is about 4 times, and the run-time is also increased by 4 times due to the additional memory usage related to the SFT layers. However, it needs to be noted that we use a single network for diverse output generation, whereas the existing methods need at least two networks for producing varying outputs. This is specifically observed in Figure 9, where it is observed that the FxSR requires less or comparable parameters than the network/image interpolation methods that use multiple ESRGAN models.

### 2.4.7 Ablation Study

The goal of classic multi-objective optimization is to find a set of solutions as close as possible to Pareto optimal front and as diverse as possible [?,?]. To investigate the performance depending on network architecture and complexity, we observe the change in the perception and distortion (PD) curve while training two versions of FxSR-PD using 16 RBs with SFT in Figure ??(a), and 23 RRDBs with SFT in Figure ??(b), respectively. As the number of training iterations increases, the PD curve of FxSR-PD converges to the desired place (lower left), and at the same time, the possible SR range on the curves is also expanded as shown in Figures ??(a) and (b). However, after a certain amount of iterations, the performance does not improve further. Figure ??(c) shows the performance comparison between the two FxSR-PD versions at the 250,000th iteration.

### 2.4.8 Discussion

#### Benefits of FxSR

A single FxSR model can produce different styles corresponding to employed feature losses and is also able to generate intermediate results between the different styles. Moreover, we can control the local regions differently by feeding a control map to the network. Hence, we can have more natural SR outputs by focusing on the foreground



or salient regions more than the backgrounds, using user-edited or automatically generated segmentation/depth/saliency maps. Also, we can remedy unnaturally generated regions by controlling the parameters as the post-processing step.

### **Limitations of FxSR**

As shown in Table 2, our method can generate comparable or superior results to the existing methods in terms of perceptual quality. But it shows a lower diversity score than the SRFlow because flat control maps are tried in this experiment. Hence, we need more studies on effective control map generation along with other feature spaces and their combinations to increase diversity.

### **Future works**

We have used a one-dimensional control parameter  $t$  for adjusting SR styles in this work. By defining more than one-dimensional SR style space with various style objectives, we can explore the  $n$ -dimensional SR spaces, possibly producing more diverse styles. Also, we may consider expanding the work to the image denoising and deblurring to control the degree of restoration locally. Furthermore, leveraging meta-learning would make it possible to improve adaptation to new samples and target objectives.

## **2.5 Conclusion**

We have presented a novel training method and a network structure for the SISR, enabling us to explore various region-wise HR outputs. From this, we can flexibly reconstruct the images between perception-oriented and distortion-oriented ones. This is achieved by defining a conditional objective function with the weights related to the perceptual losses in various feature space levels. Also, our network is designed to modulate the network’s intermediate features to change the operation according to these control inputs. As a result, we can generate an image with a desired restoration

style for each area. Experiments show that the proposed FxSR yields state-of-the-art perceptual quality and higher PSNR than other perception-oriented methods. Also, we can find many solutions by controlling a single parameter at the inference phase. We will release our code for further research and comparisons.

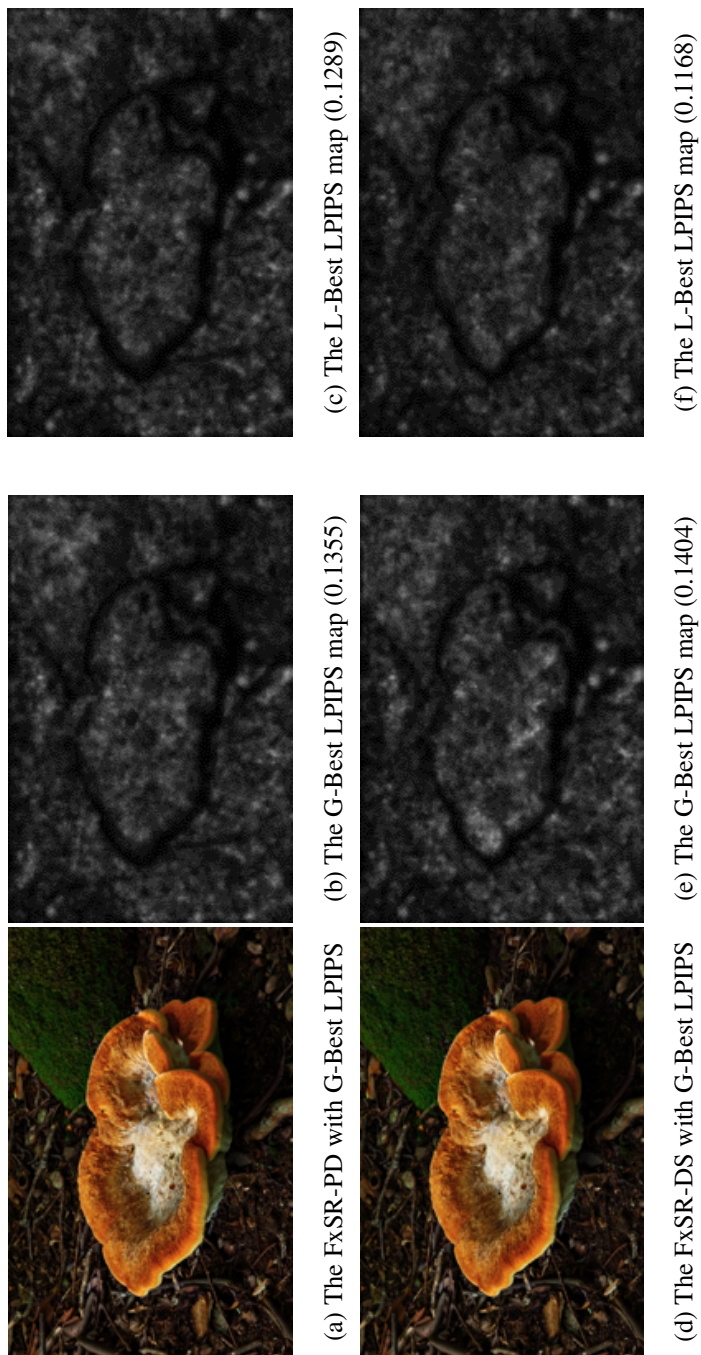
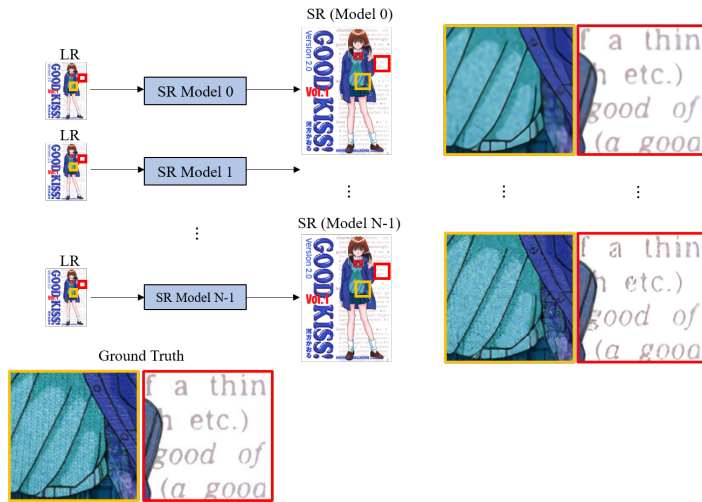
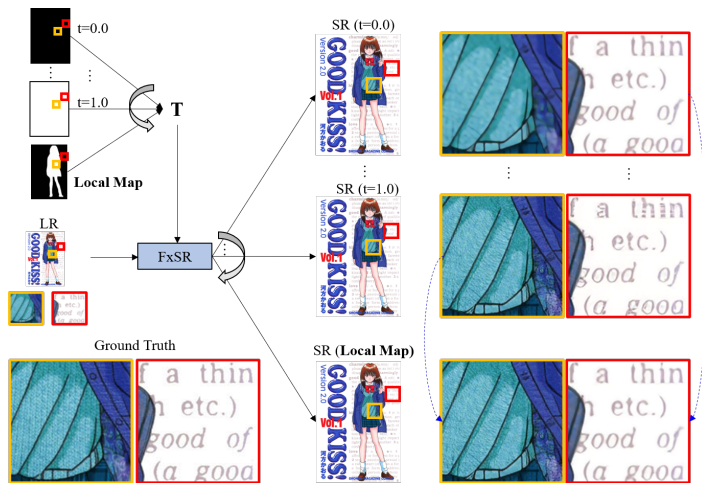


Figure 2.14: On the left are the SR results of FxSR-PD (top) and FxSR-DS (bottom) for DIV2K 0858, corresponding to t values with Global Best (G-Best) LPIPS among 11 samples, respectively. In the middle are the LPIPS maps of the SR results on the left. On the right are the Local Best (L-Best) LPIPS maps generated by selecting the highest score per pixel from 11 samples. The brighter the pixel, the higher the LPIPS value and the greater the perceptual difference from the ground truth. Each number in parentheses is the average LPIPS value for the entire image.

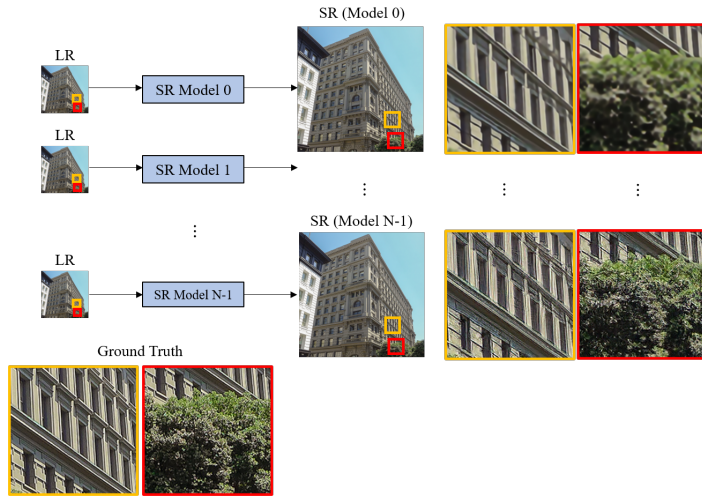


(a) The conventional method of using multiple SR models trained separately for a different objective each.

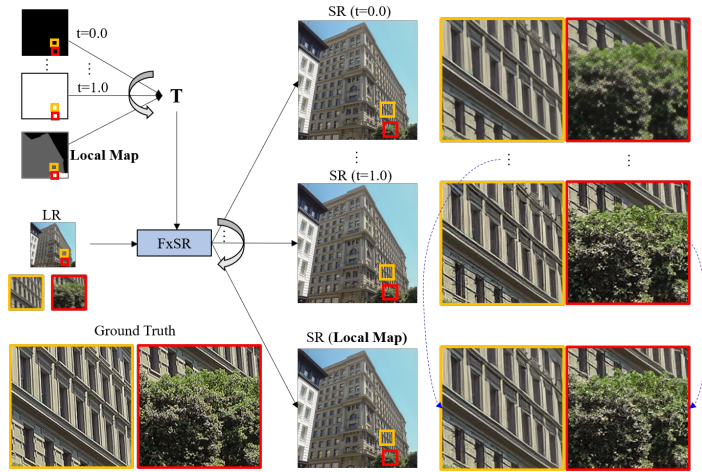


(b) The proposed method of using single FxSR-PD model trained on the training distribution of objectives.

Figure 2.15: Comparison of the SR results of the conventional method (a), which applies one objective to the entire image, and the FxSR-PD method, which applies different objectives for each area (clothes and letters) through a local map. We can see that the proposed FxSR-PD in (b) can more accurately produce the locally intended and suitable SR results without side effects such as blurry textures and broken characters.



(a) The conventional method of using multiple SR models trained separately for a different objective each.



(b) The proposed method of using single FxSR-DS model trained on the training distribution of objectives.

Figure 2.16: Comparison of the SR results of the conventional method (a), which applies one objective to the entire image, and the FxSR-DS method, which applies different objectives for each area (buildings and trees) through a local map. We can see that the proposed FxSR-DS in (b) can more accurately produce the locally intended and suitable SR results without side effects such as blurry tree textures and overshoot around the edges.

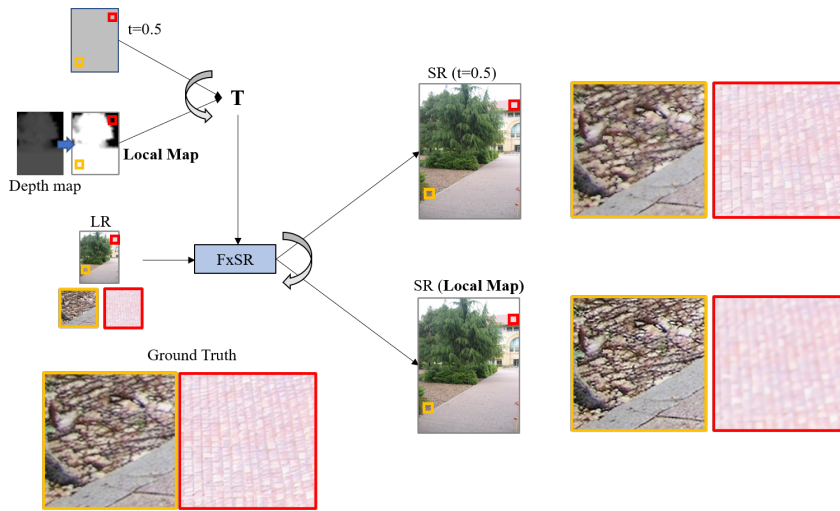


Figure 2.17: Depth-adaptive FxSR.  $T$ -maps is the modified version of the depth map of an image from the Make3D dataset [?]

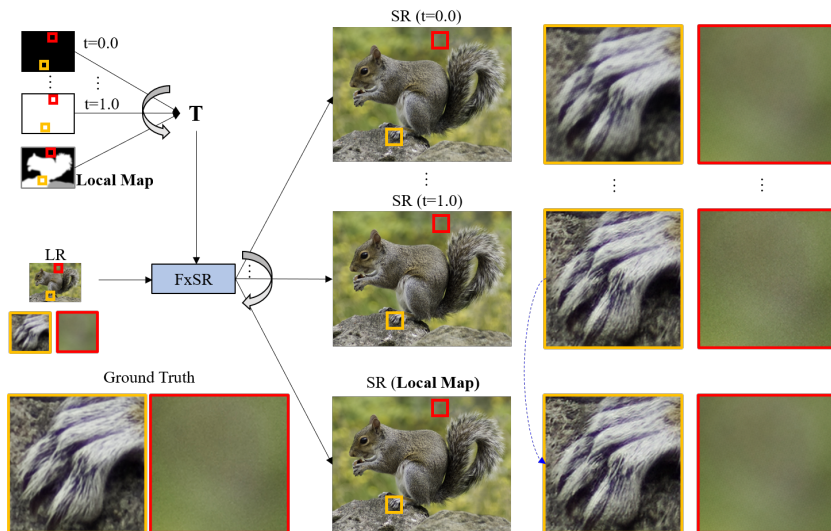
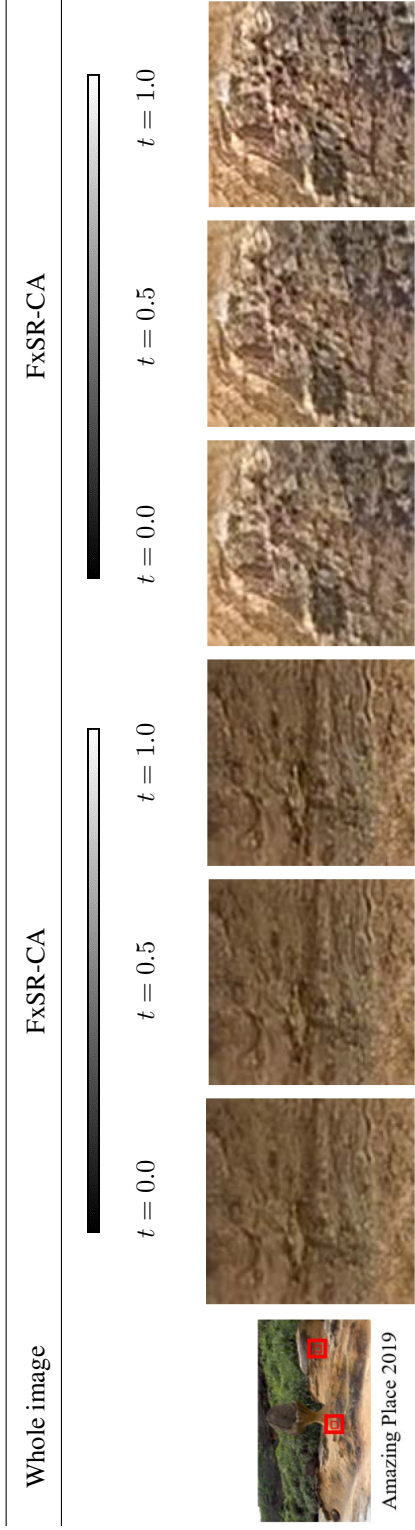


Figure 2.18: An example of applying a user-created depth map to enhance the perspective feeling with the sharper and richer textured foreground and the background with more reduced camera noise than the ground truth.



(a) An example of applying a user-created depth map to enhance the perspective feeling with the sharper and richer textured foreground and the background with more reduced camera noise than the ground truth.



(b) The intensity and style of textures change according to  $t$ .

Figure 2.19: The SR Results for compressed LR images. Two feature space (VGG44 and VGG54) and 16 RBs with SFT are used for FxSR-CA model. LR Images are extracted from "Amazing Place" video title that is encoded by VP9 codec at 0.3Mbps.

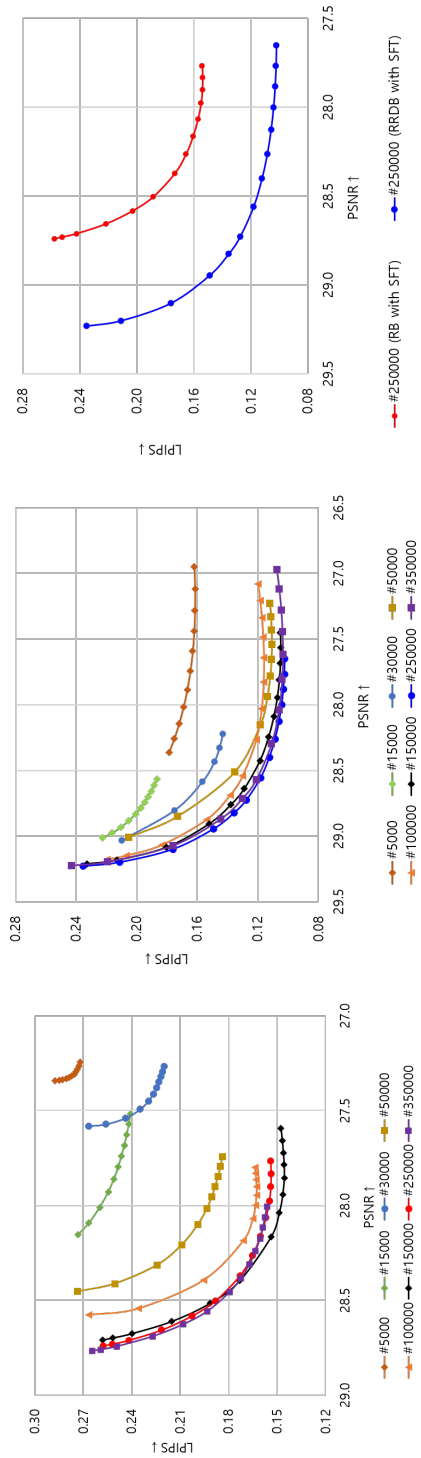


Figure 2.20: Convergence of diversity curve of the proposed FxSR-PD model as the number of training iteration increase, using (a) 16 RBs with SFT and (b) using 23 RRDBs with SFT. (c) The performance comparison between two FxSR-PD version at the 250,000th iteration



## Chapter 3

# Perception-Oriented Single Image Super-Resolution using Optimal Objective Estimation

### 3.1 Motivation and Overview

The purpose of single image super-resolution (SISR) is to estimate a high-resolution (HR) image corresponding to a given low-resolution (LR) input. SISR has many applications, mainly as a pre-processing step of computer vision or image analysis tasks, such as medical [?, ?, ?], surveillance [?, ?], and satellite image analysis [?, ?]. However, SISR is an ill-posed problem in that infinitely many HR images correspond to a single LR image. Recently, the performance of SISR has been greatly improved by adopting deep neural networks [?, ?, ?, ?, ?, ?, ?, ?, ?]. Pixel-wise distortion-oriented losses (L1 and L2) were widely used in early research, which helped to obtain a high signal-to-noise ratio (PSNR). However, these losses lead the model to generate an average of possible HR solutions, which are usually blurry and thus visually not pleasing.

Subsequently, perception-oriented losses, such as perceptual loss [?] and generative adversarial loss [?], were introduced to overcome this problem and produce realistic images with fine details [?]. Although these perception-oriented losses are used for various SR methods [?, ?, ?], they also bring undesirable side effects such as unnatural

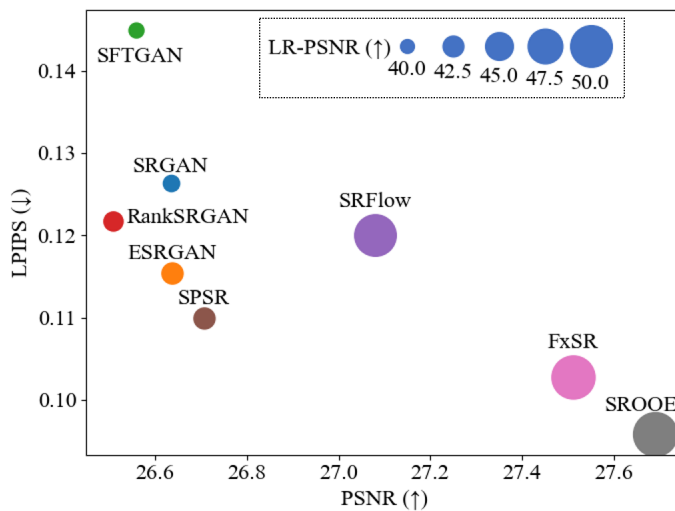
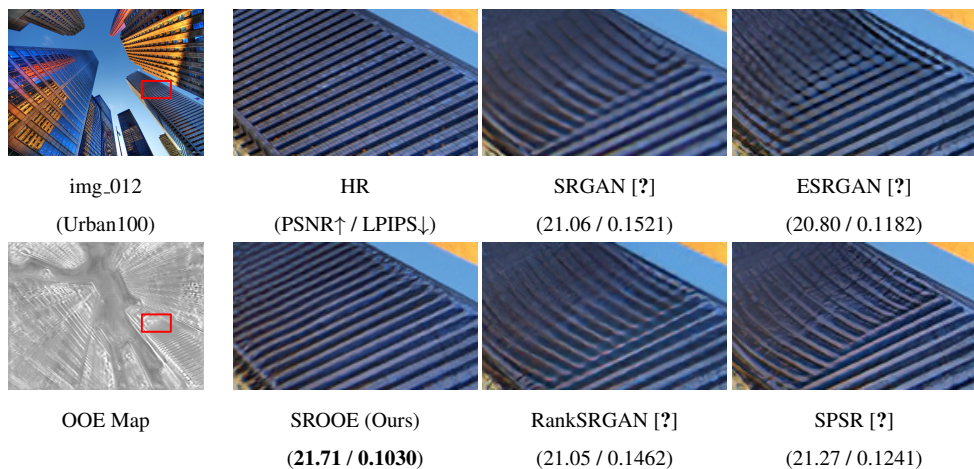


Figure 3.1: Visual and quantitative comparison. The proposed SROOE shows a higher PSNR, LR-PSNR [?] and lower LPIPS [?] than other state-of-the-art methods, *i.e.*, lower distortion and higher perceptual quality.

details and structural distortions. To alleviate these side effects and improve perceptual quality, various methods, such as the ones employing specially designed losses [?, ?] and conditional methods utilizing prior information and additional network branches [?, ?], have been introduced. Meanwhile, different from the conventional SR methods, which optimize a single objective, some studies tried to apply multiple objectives to generate more accurate HR outputs. However, some of them [?, ?, ?] applied image-specific objectives without consideration for the regional characteristics, and the other [?] used region-specific objectives for the regions obtained using semantic image segmentation with a limited number of pre-defined classes.

In this paper, we propose a new SR framework that finds a locally optimal combination of a set of objectives in the continuous sample space, resulting in regionally optimized HR reconstruction. The upper part of Fig. ?? shows a visual comparison of our results with those of state-of-the-art perception-oriented methods. We can see that our SR method using optimal objective estimation (OOE), called SROOE, generates more accurate structures. The lower part of Fig. ?? shows that the SROOE is located on the far right and bottom, corresponding to the position where both PSNR and LPIPS [?] are desirable.

For this purpose, our SR framework consists of two models: a predictive model that infers the most appropriate objectives for a given input, and a generative model that applies locally varying objectives to generate the corresponding SR result. The main challenge is to train a single generator to learn continuously varying objectives over the different locations. For this, the objective is defined as the weighted sum of several losses, and we train the generator with various sets of weights. Meanwhile, the predictor is to estimate appropriate weights for a given image input.

For efficient training, we do not learn over the entire objective space spanned by the weight vector, but find a set of several objectives that have high impacts on optimization at each vision level and are close to each other in the objective space. This is because proximity between objectives improves the efficiency of learning and in-

creases the similarity of their results, which helps reduce side effects. In addition, we train the generative model on a set of objectives on our defined trajectory, which is formed by connecting the selected objectives such that the trajectory starts with an objective suitable for a low-vision level and progresses through objectives suitable for higher levels. This enables us to replace high-dimensional weight vector manipulation with simple one-dimensional trajectory tracking, thereby simplifying the training process. The predictive model is trained using a dataset with pairs of LR images and corresponding optimal objective maps. We obtain these optimal training maps by using a grid search on the generator’s objective trajectory.

Regarding the network structure, we employ spatial feature transform (SFT) layers [?] in the generator to flexibly change the network’s behavior according to the objective. Our flexible model trained in this way has three advantages. First, the generalization capability to diversely structured images is improved since the network learns various cases. Second, the SR results are consistent with respect to the trajectory and given input. Third, the high-dimensional weight vector for loss terms can be replaced with a vector function with a one-dimensional input, and thus the optimal loss combinations can be easily found and controlled.

Our contributions are summarized as follows. (1) We propose an SISR framework that estimates and applies an optimal combination of objectives for each input region and thus produces perceptually accurate SR results. (2) While this approach requires training with various weighted combinations of losses, which needs the search on a high-dimensional weight vector space, we introduce an efficient method for exploring and selecting objectives by defining the objective trajectory controlled by a one-dimensional variable. (3) We propose a method for obtaining optimal objective maps over the trajectory, which are then used to train the objective estimator. (4) Experiments show that our method provides both high PSNR and low LPIPS, which has been considered a trade-off relation.

## 3.2 Related Work

**Distortion-oriented SR.** Dong *et al.* [?] first proposed a convolutional neural network (CNN)-based SR method that uses a three-layer CNN to learn the mapping from LR to HR. Since then, many deeper CNN-based SISR frameworks have been proposed [?, ?]. Ledig *et al.* [?] proposed SRResNet, which uses residual blocks and skip-connections to further enhance SR results. Since Huang *et al.* [?] proposed DenseNet, the dense connections have become prevalent in SR networks [?, ?, ?, ?, ?]. Zhang *et al.* [?] introduced RCAN, which employs channel attention and improves the representation ability of the model and SR performance. More recently, SwinIR [?] and Uformer [?] reported excellent SISR performance by using the Swin Transformer architecture [?] and locally-enhanced window (LeWin) Transformer block, respectively. While there are many architectures for the SR as listed above, we employ plain CNN architectures as our predictor and generator. The structure is not an issue in this paper, and various CNNs and Transformers can be tried instead of our architecture.

**Perception-oriented SR.** Because the pixel losses, such as L1 and L2, do not consider perceptual quality, the results of using such losses often lack high-frequency details [?, ?]. Meanwhile, Johnson *et al.* [?] proposed a perceptual loss to improve the visual quality of the output. Ledig [?] introduced SRGAN utilizing adversarial loss [?], which can generate photo-realistic HR images. Wang *et al.* [?] enhanced this framework by introducing ESRGAN with Residual-in-Residual Dense Block (RRDB).

However, these perception-oriented SR models entail undesirable artifacts, such as unexpected textures on a flat surface. To alleviate such artifacts and/or further improve the perceptual quality, various methods have been proposed. Soh *et al.* [?] introduced NatSR, where they designed a loss to suppress aliasing. Wang *et al.* [?] proposed the use of semantic priors for generating semantic-specific details by using SFT layers. Zhang *et al.* [?] proposed a Ranker that learns the behavior of perceptual metrics. Ma *et al.* [?] proposed a structure-preserving super-resolution (SPSR) to alleviate geometric distortions. Liang *et al.* [?] proposed locally discriminative learning be-

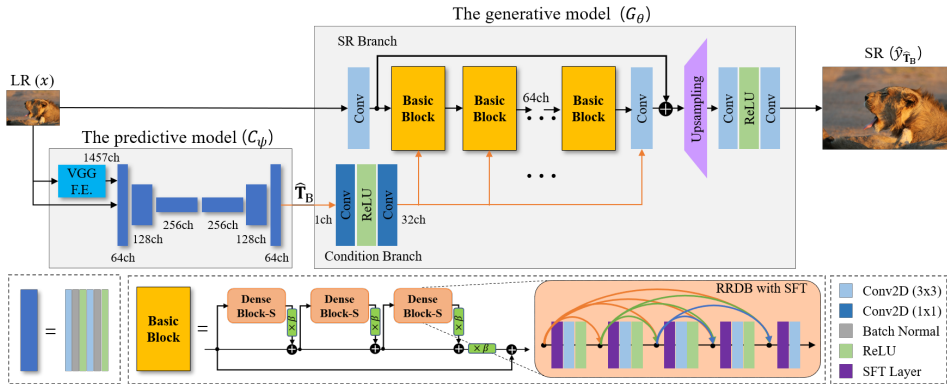


Figure 3.2: Architecture of the proposed method. The predictive model generates the optimal objective map  $\hat{\mathbf{T}}_B$ , which is fed to the generative model. The input LR image is super-resolved through our Basic Blocks and other elements of the generator, which are controlled by the map from the Condition Branch.

tween GAN-generated artifacts and realistic details. However, Blau [?] argued that it is difficult to simultaneously achieve perceptual quality enhancement and distortion reduction because they involve a trade-off relationship. In this regard, there was an SR challenge [?] focused on the trade-off between generation accuracy and perceptual quality. One of the main claims of this paper is that we can further reduce distortion and increase perceptual quality simultaneously, as shown in Fig. ??.

### 3.3 Methods

#### 3.3.1 Proposed SISR Framework

An overview of our SISR framework is presented in Fig. ?? . Our framework consists of a predictive model  $C_\psi$  and generative model  $G_\theta$ , parameterized by  $\psi$  and  $\theta$ , respectively. Model  $C_\psi$  infers an LR-sized optimal objective map  $\hat{\mathbf{T}}_B$  for a given LR input  $x$ , and  $G_\theta$  applies it to produce the corresponding SR output, which is as similar as

possible to its corresponding HR counterpart  $y$ , as follows:

$$\hat{y}_{\hat{\mathbf{T}}_B} = G_\theta \left( x | \hat{\mathbf{T}}_B \right), \quad (3.1)$$

$$\hat{\mathbf{T}}_B = C_\psi(x). \quad (3.2)$$

### 3.3.2 Proposed Generative Model

Since using a single fixed objective cannot generate optimized HR results for every image region, it is beneficial to apply regionally different losses regarding the input characteristics. However, training multiple SR models, each of which is trained with a different objective, is impractical because it requires large memory and long training and inference times [?]. Hence, in this paper, we propose a method to train a single SR model that can consider locally different objectives.

**Effective Objective Set.** We first investigate which objectives need to be learned for accurate SR. For perception-oriented SR [?, ?], the objective is usually a weighted sum of pixel-wise reconstruction loss  $\mathcal{L}_{rec}$ , adversarial loss  $\mathcal{L}_{adv}$ , and perceptual loss  $\mathcal{L}_{per}$ , as follows:

$$\mathcal{L} = \lambda_{rec} \cdot \mathcal{L}_{rec} + \lambda_{adv} \cdot \mathcal{L}_{adv} + \sum_{per_l} \lambda_{per_l} \cdot L_{per_l}, \quad (3.3)$$

$$L_{per_l} = \mathbb{E} [\|\phi_{per_l}(\hat{y}) - \phi_{per_l}(y)\|_1], \quad (3.4)$$

$$per_l \in \{\text{V12, V22, V34, V44, V54}\}, \quad (3.5)$$

where  $\lambda_{rec}$ ,  $\lambda_{adv}$ , and  $\lambda_{per_l}$  are weighting parameters for the corresponding losses, and  $\phi_{per_l}(\cdot)$  represents feature maps of the input extracted at layer  $per_l$  of the 19-layer VGG network, where five layers denoted in Eqn. ?? are considered as in [?, ?, ?]. Since the receptive field becomes larger as we progress deeper into the VGG network [?], features of shallow layers such as V12 and V22 and deeper layers such as V34, V44, and V54 correspond to relatively low-level and higher-level vision, respectively [?].

To find an effective set of objectives, we define an SR objective space. Since the objective for SR is a weighted sum of seven loss terms, as in Eqn. ??, an objective space

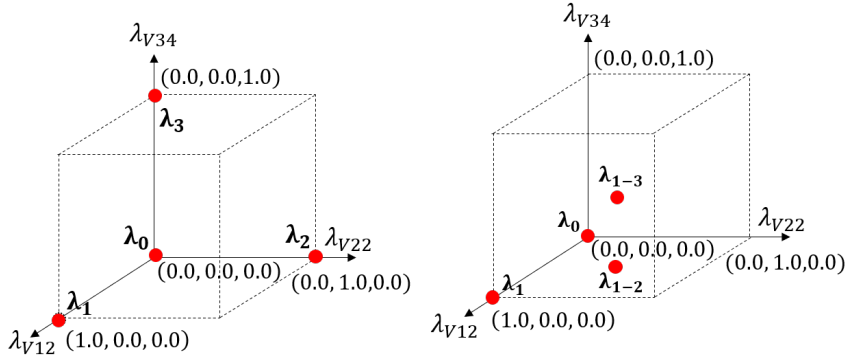


Figure 3.3: Set  $A$  (left) and set  $B$  (right) in the objective space. The objectives in set  $B$  are closer each other than those in set  $A$ .

is spanned by these basis loss terms, and any objective can be expressed by a seven-dimensional vector of weighting parameters,  $\lambda_i \in \mathbb{R}^7$  as  $\lambda_i = [\lambda_{rec}, \lambda_{adv}, \lambda_{per}]$ , where  $\lambda_{per} \in \mathbb{R}^5$  is a weight vector for perceptual loss.

Table ?? compares two objective sets,  $A$  and  $B$ , defined as shown in Fig. ?. Because ESRGAN [?] is the base model for this comparison, for all objectives in the table, except for  $\lambda_0$ ,  $\lambda_{rec}$  and  $\lambda_{adv}$  are set to  $1 \times 10^{-2}$  and  $5 \times 10^{-3}$ , respectively. These are the same as those for ESRGAN, except that  $\lambda_{per}$  changes, where  $\|\lambda_{per}\|_1 = 1$ . In particular, in terms of  $\lambda_{per}$ , whereas each objective  $\lambda_i$  in set  $A$  has weights for only one of the five VGG feature spaces, each objective in set  $B$  has equal weights for each loss in the feature space lower than the target vision level. Therefore, an objective corresponding to a high vision level also includes the losses for the lower-level feature spaces. Meanwhile, because  $\lambda_0$  corresponds to a distortion-oriented RRDB model [?], its  $\lambda_{rec}$  and  $\lambda_{adv}$  are set to  $1 \times 10^{-2}$  and 0, respectively. Note that  $\lambda_0$  is included in both sets  $A$  and  $B$ .

In Table ??, the normalized versions (min-max feature scaling) of the averaged  $L_{per_l}$  from Eqn. ?? for five datasets (BSD100 [?], General100 [?], Urban100 [?], Manga109 [?], and DIV2K [?]) are reported. For all feature spaces, including the targeted V12 and V22 feature spaces,  $\lambda_{1-2}$  in set  $B$  has smaller L1 errors than those of  $\lambda_1$  and  $\lambda_2$  in set  $A$ . Moreover,  $\lambda_{1-4}$  exhibits smaller errors than those of  $\lambda_4$  and  $\lambda_5$ .



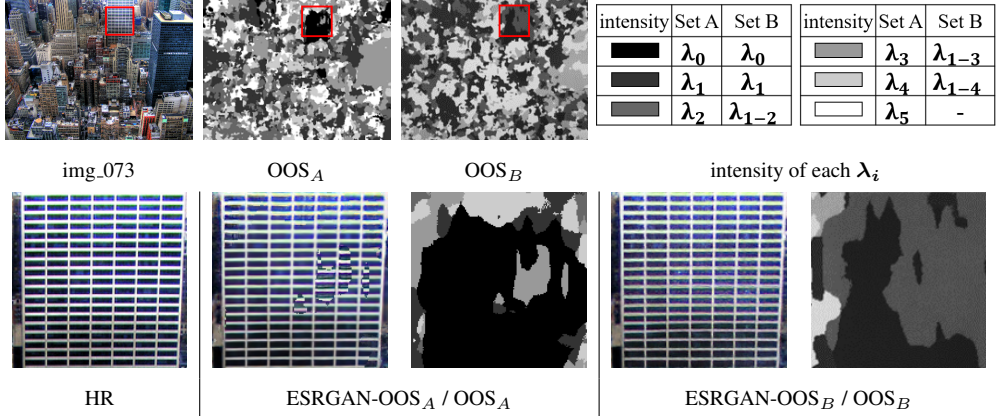


Figure 3.4: The  $OOS_A$  and  $OOS_B$  results using Sets  $A$  and  $B$  (top), their SR results,  $ESRGAN-OOS_A$  and  $ESRGAN-OOS_B$  (bottom).

Although  $\lambda_{1-3}$  has slightly more errors in the V34 feature space than that of  $\lambda_3$ , it has less errors in the V12 and V22 feature spaces therefore,  $\lambda_{1-3}$  has relatively less distortion than  $\lambda_3$  overfitted to the V34 feature space. That is supported by the fact that most of the objectives in set  $B$ , including  $\lambda_{1-3}$ , have better PSNR and LPIPS on Urban100 [?] than those in set  $A$ .  $\lambda_{1-5}$  showing relatively poor performance compared to  $\lambda_{1-4}$  is not used.

To examine the SR result with locally appropriate objectives applied using set  $A$ , we mix the six SR results of  $ESRGAN-\lambda_a$ , where  $\lambda_a \in A$ , by selecting the SR result with the lowest LPIPS for each pixel position, as follows:

$$y_A^*(i, j) = \hat{y}_{\mathbf{T}_A^*(i, j)}(i, j), \quad (3.6)$$

$$\mathbf{T}_A^*(i, j) = \arg \min_{\lambda_a \in A} \mathbf{LPIPS}_{\lambda_a}(i, j), \quad (3.7)$$

$$\mathbf{LPIPS}_{\lambda_a} = \mathbf{LPIPS}(y, \hat{y}_{\lambda_a}), \quad (3.8)$$

where  $\hat{y}_{\lambda_a}$  is the SR result of  $ESRGAN-\lambda_a$ . The  $LPIPS$  function computes the perceptual distance between two image patches for each pixel position, producing an LPIPS map,  $\mathbf{LPIPS}$ , of the input image size [?, ?]. The LPIPS metric in Table ?? is the average of this map. Since  $\mathbf{T}_A^*$  is the optimal objective selection (OOS),  $\mathbf{T}_A^*$  and

its SR model for mixing are denoted as  $OOS_A$  and  $ESRGAN-OOS_A$ , respectively. The upper part of Fig. ?? shows an example of  $OOS_A$  and  $OOS_B$  based on set  $A$  and  $B$ . PSNR and LPIPS [?] of  $ESRGAN-OOS_A$  and  $ESRGAN-OOS_B$  are reported in Table ??, where  $ESRGAN-OOS_B$  is superior to any single objective model, demonstrating the potential for performance improvement of the locally suitable objective application. The lower part of Fig. ?? shows the side effects caused by mixing the SR results for set  $A$  with lower proximities between objectives than those in set  $B$ , as shown in Fig. ?. Since  $ESRGAN-OOS_B$  in Fig. ? has less artifact and better PSNR than those of  $ESRGAN-OOS_A$ , the proposed set  $B$  is more suitable for applying locally appropriate objectives.

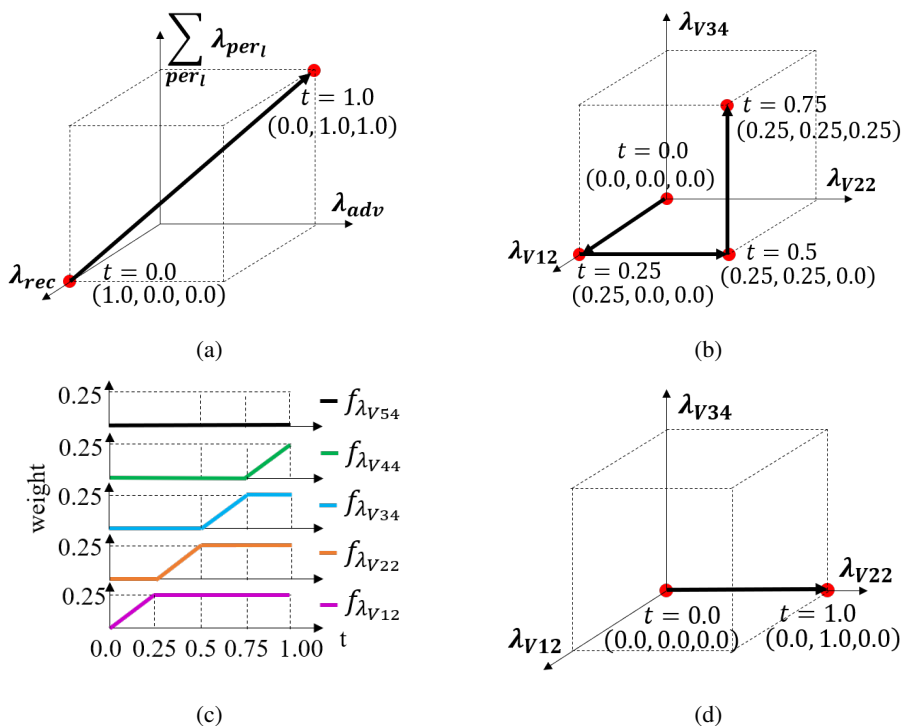


Figure 3.5: The proposed vector functions for loss weights, (a)  $\lambda(t)$  in Eqn. ?? when  $\alpha=1$  and  $\beta=0$ , (b) its  $\lambda_{per}(t)$  and (c) the weighting functions for  $\lambda_{per}(t)$ . (d)  $\lambda_{per}(t)$  used for FxSR [?].

**Learning Objective Trajectory.** We train our generative model on a set of ob-

jectives over the trajectory rather than a single objective,  $\lambda_i$ . The objective trajectory is formed by connecting the selected objectives, i.e. the five objectives of set  $B$ , starting with an objective for a low-vision level and progressing through objectives for higher levels, i.e. from  $\lambda_0$  to  $\lambda_{1-4}$ . It is parameterized by a single variable  $t$ ,  $\lambda(t) = \langle \lambda_{rec}(t), \lambda_{adv}(t), \lambda_{per}(t) \rangle$ , as follow:

$$\lambda(t) = \alpha \cdot f_\lambda(t) + \beta, \quad (3.9)$$

$$f_\lambda(t) = \langle f_{\lambda_{rec}}(t), f_{\lambda_{adv}}(t), f_{\lambda_{per}}(t) \rangle, \quad (3.10)$$

where  $f_{\lambda_{per}}(t) \in \mathbb{R}^5$ ,  $f_{\lambda_{rec}}(t)$ ,  $f_{\lambda_{adv}}(t)$  are weighting functions,  $\alpha$  and  $\beta$  are the scaling and offset vectors. As  $f_\lambda : \mathbb{R} \rightarrow \mathbb{R}^7$ , this vector function enables the replacement of high-dimensional weight-vector manipulation with one-dimensional tracking, simplifying the training process.

Specifically, the trajectory design is based on the observation in Table ?? that the distortion-oriented RRDB model using  $\lambda_0$  has smaller L1 errors than those of all ESRGAN models for low-level feature spaces, such as V12 and V22, whereas ESRGAN models have smaller L1 errors for higher-level feature spaces, such as V34, V44, and V54. Thus, we design the weight functions  $f_{\lambda_{rec}}$ ,  $f_{\lambda_{adv}}$  and  $f_{\lambda_{per}}$  such that when  $t$  approaches 0,  $f_{\lambda_{rec}}$  increases and  $\{f_{\lambda_{adv}}, \sum_{per_l} f_{\lambda_{per_l}}\}$  decrease to go to  $\lambda_0$ , and conversely to go to  $\lambda_{1-4}$  when  $t$  increases to 1, as shown in Fig. ??(a).

In relation to the change in  $\sum_{per_l} f_{\lambda_{per_l}}(t)$ , we design each of five component functions,  $f_{\lambda_{per_l}}(t)$  of  $f_{\lambda_{per}}(t)$ , as shown in Fig. ??(c), to obtain the objective trajectory from  $\lambda_0$  to  $\lambda_{1-4}$  of set  $B$  as shown in Fig. ??(b), illustrating only three out of five components because of the limitations of 3-dimensional visualization. Thus, as we progress through the trajectory by increasing  $t$  from 0 to 1, the weighting parameters for the objective start with the distortion-oriented objective,  $\lambda_0$ , and then the losses of higher-vision-level feature spaces and adversarial loss are progressively added, making slight transitions on the objective toward  $\lambda_{1-4}$ . Fig. ??(d) shows the objective trajectory used for FxSR [?], which uses only the V22 feature space, limiting

the performance of the perceptually accurate restoration.

The proposed objective trajectory can efficiently improve the accuracy and consistency of the SR results. First, we can use any objective on the continuous trajectory from low to high-level vision, which allows the application of more accurate objectives to each region. Second, with regard to consistency, high-level objectives on our proposed trajectory include both low-level and high-level losses, thus also accounting for the low-level objectives. This weighting method allows the sharing of the structural components reconstructed mainly by low-vision-level objectives between all SR results on the trajectory. Finally, we need to train a single SR model only once, reducing the number of models required to produce diverse HR outputs [?, ?].

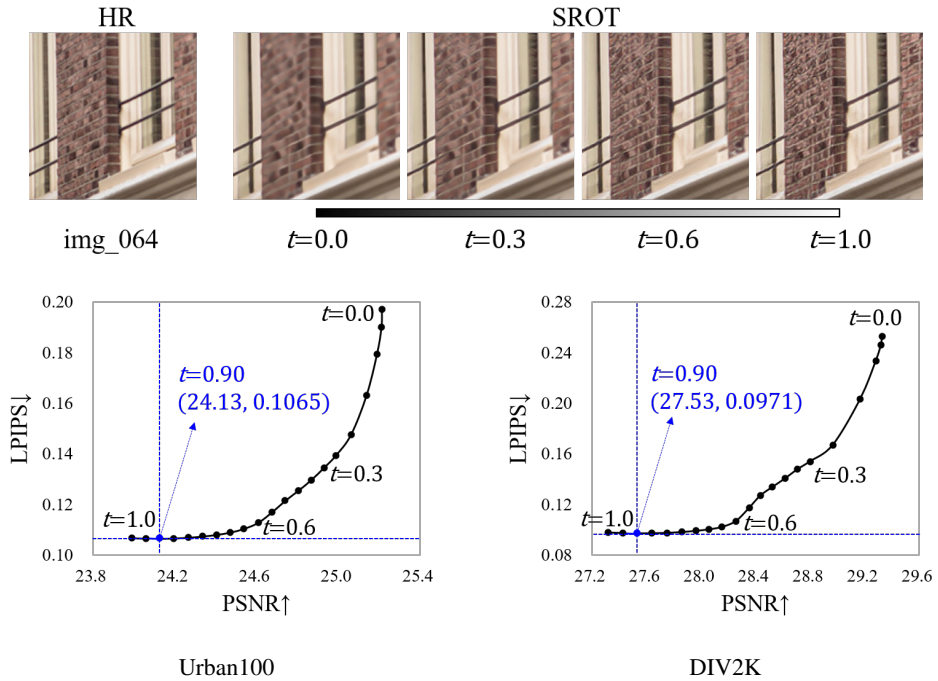


Figure 3.6: Changes in detail in the SROT results according to  $t$ -value (top) and changes in PSNR and LPIPS for test DBs (bottom).

Fig. ?? shows the changes in the result of the generative SR model trained on the objective trajectory in Fig. ??(b), called SROT, as  $t$  changes from 0 to 1. The graphs

in the bottom of Fig. ?? shows the trade-off curves in the perception-distortion plane according to the change of  $t$ , where  $t$  increases by 0.05 from 0.0 to 1.0 and has 21 sample points. Each SR result on the curve is obtained by inputting  $\mathbf{T}$  with the same  $t$  throughout the image, as  $\mathbf{T}_t = \mathbf{1} \times t$ , into the condition branch of the generative model, as follows:

$$\hat{y}_{\mathbf{T}_t} = G_\theta(x|\mathbf{T}_t). \quad (3.11)$$

The horizontal and vertical dotted lines of the graphs in Fig. ?? indicate the lowest LPIPS values of the model and the corresponding PSNR values, respectively. The  $t$  values at that time are written next to the vertical lines. However, applying a specific  $t$  to the entire image still limits SR performance, and optimal  $t$  depending on images is unknown at inference time. We take this one step further and present later how to estimate and apply locally optimal objectives.

**Network Architecture and Training.** The outline of the generator network is adopted from [?], *i.e.*,  $G_\theta$  consists of two streams, an SR branch with 23 basic blocks and a condition branch as shown in Fig. ?. The condition branch takes an LR-sized target objective map  $\mathbf{T}$  and produces shared intermediate conditions that can be transferred to all the SFT layers in the SR branch. Since the SFT layers [?] modulate feature maps by applying affine transformation, they learn a mapping function that outputs a modulation parameter based on  $\mathbf{T}$ . Specifically,  $\mathbf{T}_t$ , with  $t$  randomly changing in the pre-defined range, is fed into the condition branch during training, and this modulation layer allows the SR branch to optimize the changing objective by  $t$ . As a result,  $G_\theta$  learns all the objectives on the trajectory and generates SR results with spatially different objectives according to the map at inference time.  $G_\theta$  is optimized on the training samples  $\mathcal{Z} = (x, y)$  with the distribution  $P_{\mathcal{Z}}$ , as follows:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{\mathcal{Z} \sim P_{\mathcal{Z}}} [\mathcal{L}(\hat{y}_{\mathbf{T}_t}, y|t)], \quad (3.12)$$

$$\mathcal{L}(t) = \lambda_{rec}(t) \cdot \mathcal{L}_{rec} + \lambda_{adv}(t) \cdot \mathcal{L}_{adv} + \sum_{per_l} \lambda_{per_l}(t) \cdot L_{per_l}. \quad (3.13)$$

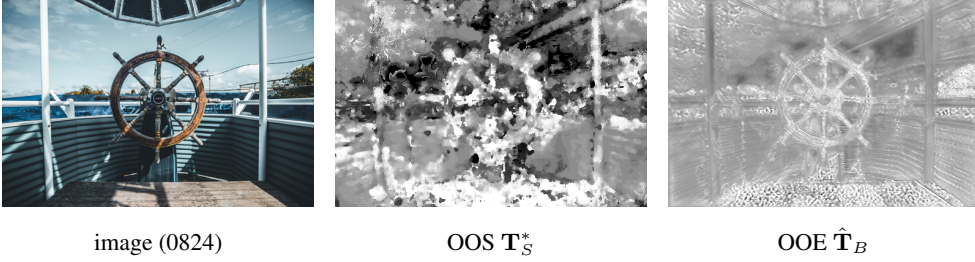


Figure 3.7: The input image, the optimal objective selection  $\mathbf{T}_S^*$  obtained by parameter sweeping, and  $\hat{\mathbf{T}}_B$  estimated by  $C_\psi$ .

### 3.3.3 Optimal Objective Estimation (OOE)

To estimate an optimal combination of objectives for each region, we train a predictive model,  $C_\psi$ . This model produces an optimal objective map  $\hat{\mathbf{T}}_B$  estimated for a given LR image, which is then delivered to the generative model in Eqn. ???. Since it is hard to find the ground truth map for  $C_\psi$  training, we obtain its approximation  $\mathbf{T}_S^*$  via a simple exhaustive searching to narrow down the range of the best possible values. Specifically, we generate a set of 21 SR results by changing  $t$  from 0 to 1 with a step of 0.05, and the optimal objective maps are generated by selecting the  $t$  with the lowest LPIPS among them for each pixel, as

$$\mathbf{T}_S^*(i, j) = \arg \min_{t \in S} \text{LPIPS}_t(i, j), \quad (3.14)$$

$$\text{LPIPS}_t = \text{LPIPS}(y, \hat{y}_{\mathbf{T}_t}), \quad (3.15)$$

where  $\mathbf{T}_t = \mathbf{1} \times t, t \in S = \{0.0, 0.05, 0.10, \dots, 1.0\}$ . Fig. ?? shows an example of the optimal objective selection (OOS)  $\mathbf{T}_S^*$ . The SR result using  $\mathbf{T}_S^*$ , SROOS, can be an upper-bound approximation for the performance of  $G_\theta$  as

$$\hat{y}_{\mathbf{T}_S^*} = G_\theta(x | \mathbf{T}_S^*). \quad (3.16)$$

Although  $\mathbf{T}_S^*$  is useful for training  $C_\psi$ , this pixel-wise objective selection without considering the interference caused by the convolutions of  $G_\theta$  is not accurate ground truth. Therefore,  $C_\psi$  is optimized with three loss terms: pixel-wise objective map loss,













					
0841 (DIV2K)	HR	RRDB	SROOE ( $\mathbf{T} = \mathbf{0}$ )	SRGAN	ESRGAN
	(PSNR $\uparrow$ / SSIM $\uparrow$ / LPIPS $\downarrow$ )	(28.73 / 0.8851 / 0.1929)	(28.60 / 0.8813 / 0.1927)	(25.89 / 0.8128 / 0.1141)	(25.98 / 0.8182 / 0.1048)
					
OOE Map ( $\hat{\mathbf{T}}_B$ )	SFTGAN	RankSRGAN	SRFlow	SPSR	SROOE
	(25.76 / 0.8004 / 0.1370)	(25.83 / 0.8046 / 0.1098)	(26.45 / 0.7963 / 0.1093)	(26.32 / 0.8182 / 0.1076)	<b>(26.58 / 0.8283 / 0.0897)</b>

Figure 3.8: Visual comparison with state-of-the-art SR methods. Among the seven perception-oriented SR methods, the best performances are highlighted in **bold**.

pixel-wise reconstruction loss and perceptual loss, which measures the difference between the reconstructed and HR images, as follows:

$$\psi^* = \arg \min_{\psi} \mathbb{E}_{\mathcal{Z}_{\mathbf{T}} \sim \mathcal{P}_{\mathcal{Z}_{\mathbf{T}}}} \mathcal{L}, \quad (3.17)$$

$$\mathcal{L} = \lambda_{\mathbf{T}} \cdot \mathcal{L}_{\mathbf{T}} + \lambda_{rec}^{OOE} \cdot \mathcal{L}_{rec} + \lambda_R \cdot \mathcal{L}_R, \quad (3.18)$$

$$\mathcal{L}_R = \mathbb{E} \left[ LPIPS \left( y, \hat{y}_{\hat{\mathbf{T}}_B} \right) \right], \quad (3.19)$$

where  $\mathcal{L}_{\mathbf{T}}$  and  $\mathcal{L}_{rec}$  is the L1 losses between  $\mathbf{T}_S^*$  and  $\hat{\mathbf{T}}_B$  and between  $y$  and  $\hat{y}_{\hat{\mathbf{T}}_B}$ , respectively. Meanwhile,  $\mathcal{Z}_{\mathbf{T}} = (x, y, \mathbf{T}_S^*)$  is the training dataset, and  $\lambda_{\mathbf{T}}$ ,  $\lambda_{rec}^{OOE}$  and  $\lambda_R$  are the weights for each of the loss terms, respectively. During the  $C_{\psi}$  model training,  $C_{\psi}$  is combined with the already trained generative model, and the generator parameters are fixed. Therefore, losses for  $C_{\psi}$  training, including LPIPS, are involved only in estimating locally-appropriate objective maps without changing the generator’s parameters.

The architecture of  $C_{\psi}$  consists of two separate sub-network: one is a feature extractor (F.E.) utilizing the VGG-19 [?] and the other is a predictor with the UNet ar-

chitecture [?], as shown in Fig ?? . For better performance, the feature extractor aims to get low to high-level features and delivers them to Unet, which makes the prediction. Since the structure of UNet has a wider receptive field, it is advantageous for predicting objectives in context.

## 3.4 Experiments

### 3.4.1 Experiment Setup

**Materials, Evaluation Metrics and Training Details.** We use either the DIV2K [?] (800 images) or the DF2K [?] (3450 images) dataset to train our models. Our test datasets include BSD100 [?], General100 [?], Urban100 [?], Manga109 [?], and DIV2K validation set [?]. To evaluate the perceptual quality, we report LPIPS [?] and DISTS [?], which are full-reference metrics. DISTS is a perceptual metric that focuses on detail similarity. PSNR and SSIM [?] are also reported as fidelity-oriented metrics. The LR-PSNR metric is the PSNR between the LR input and downscaled SR images. The higher the LR-PSNR, the better the consistency between the SR results and LR images, where 45 dB or more is recommended for good LR consistency as addressed in NTIRE challenge [?]. Because consistency with the LR input images is important, we also report the LR-PSNR. All training parameters are set to be equal to those of ESRGAN [?], except for the loss weights. For the generator training,  $t$  is a random variable with uniform distribution in  $[0, 1]$ .  $\alpha=[1 \times 10^{-2}, 1, 1]$  and  $\beta=[1 \times 10^{-2}, 0, 0]$ .

### 3.4.2 Evaluation

**Quantitative Comparison.** Table ?? shows the quantitative performance comparison for the  $4\times$  SR. We compared it with a distortion-oriented method, RRDB [?], and perception-oriented methods, such as SRGAN [?], ESRGAN [?], SFTGAN [?], RankSRGAN [?], SRFlow [?], SPSR [?], and FxSR [?]. The table shows that our method yields the best results among the perception-oriented methods on all datasets,



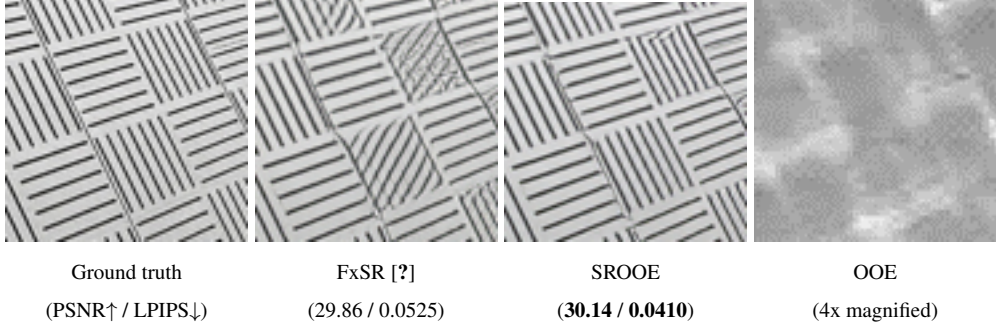


Figure 3.9: Visual comparison of the results of FxSR and SROOE.

not only in terms of LPIPS [?] and DISTs [?], but also in terms of distortion-oriented metrics such as PSNR and SSIM. It also exceeds 45 dB in LR-PSNR, indicating that LR consistency is well-maintained, as addressed in NTIRE [?]. In addition, SROOE using a local objective map outperforms SROT with the globally optimal  $t$  value for the Urban100 and DIV2K benchmarks in terms of both LPIPS and PSNR in Fig. ???. SROOS with  $\mathbf{T}_\zeta^*$  has the best PSNR, SSIM, LPIPS, and DISTs scores, which shows the approximated upper bounds of the proposed SROOE. On the other hand, when the objective map  $\mathbf{T}$  is set to be  $\mathbf{0}$ , SROOE operates as a distortion-oriented SR model. Although it is slightly inferior to RRDB [?] in terms of PSNR, its performance is not far behind while showing better LPIPS. This implies that SROOE performs close to RRDB [?] for the regions needing distortion-oriented restoration, and thus the overall distortion is reduced while achieving high perceptual quality.

**Qualitative Comparison.** Fig. ?? shows a visual comparison, where we can observe that SROOE generates more accurate structures and details. In particular, it appears that there is little change in the structural component between the SROOE results using  $\mathbf{T} = \mathbf{0}$  and  $\hat{\mathbf{T}}_B$ , and sharp edges and generated details are added to the structural components. Additional visual and quantitative comparisons for the  $4\times$  and  $8\times$  SR are provided in the supplementary.

### 3.5 Ablation Study

Table ?? reports average values in terms of each metric on all five benchmarks in Table ??, which vary according to the change in each element. The two different objective trajectories shown in Fig. ??(b) and (d) are referred to as P1234 and P2, respectively. The table shows that the SR performance improves step by step, when going from P2, fixed  $t$ , and DIV2K to P1234, OOE, and DF2K, respectively. Specifically, our proposed model, SROOE-P1234 trained with DF2K is improved by 0.25 dB in PSNR, 0.0069 in SSIM, 0.0051 in LPIPS, and 0.23 dB in LR-PSNR compared to SROT-P2 corresponding to FxSR with  $t=0.8$  [?] as the example shown in Fig. ?. A comparison of the running times and parameter sizes is presented in Table ??, where the time is for the  $4\times$  SR of a  $128 \times 128$  image on an NVIDIA RTX3090 GPU.

**Limitations.** Although applying locally appropriate objectives can significantly improve the LR to HR mapping accuracy, even if the generator uses an optimal objective map  $\mathbf{T}_S^*$ , it is still limited in achieving full reconstruction. This means that the proposed generator is still unable to generate all HRs with the objective set used for training in this study, and thus, more sophisticated perceptual loss terms than the VGG feature space are still required to overcome this. And still, there remains a limit to solving the ill-posed problem caused by high-frequency loss.

### 3.6 Conclusion

We proposed a novel SISR framework for perceptually accurate HR restoration, where an objective estimator provides an optimal combination of objectives for a given image patch, and a generator produces SR results that reflect the target objectives. For this purpose, we employed objective trajectory learning to efficiently train a single generative model that can apply varying objective sets. Experiments show that the proposed method reduces visual artifacts, such as structural distortion and unnatural details, and achieves improved results compared to those of state-of-the-art perception-oriented

methods in terms of both perceptual and distortion metrics. The proposed method can be applied to off-the-shelf and other SISR network architectures.

Table 3.1: Performance comparison of SR results of ESRGAN models with different weight vectors for perceptual loss. Among the objectives in Sets  $A$  and  $B$ , except for  $\lambda_0$ , the 1st and the 2nd best performances for each column are highlighted in **bold** and underline.

Set	objec- tive	$\lambda_{per}$ [ $\lambda_{V12}, \lambda_{V22}, \lambda_{V34}, \lambda_{V44}, \lambda_{V54}$ ]	Normalized $L_{per1}$				Metric		
			$L_{V12}$	$L_{V22}$	$L_{V34}$	$L_{V44}$	$L_{V54}$	PSNR	LPIPS
	$\lambda_0$	[0.0, 0.0, 0.0, 0.0, 0.0]	0.00	0.00	1.00	1.00	1.00	25.48	0.1960
$A$	$\lambda_1$	[1.0, 0.0, 0.0, 0.0, 0.0]	<u>0.71</u>	0.53	0.76	0.38	0.25	<u>23.95</u>	0.1124
	$\lambda_2$	[0.0, 1.0, 0.0, 0.0, 0.0]	0.72	<u>0.26</u>	0.35	0.22	0.15	23.84	0.1125
	$\lambda_3$	[0.0, 0.0, 1.0, 0.0, 0.0]	0.81	0.51	<b>0.00</b>	<b>0.02</b>	0.04	23.66	0.1124
	$\lambda_4$	[0.0, 0.0, 0.0, 1.0, 0.0]	0.92	0.78	0.51	0.11	0.05	23.28	0.1158
	$\lambda_5$	[0.0, 0.0, 0.0, 0.0, 1.0]	1.00	1.00	0.93	0.32	0.12	23.00	0.1232
$B$	$\lambda_1$	[1.0, 0.0, 0.0, 0.0, 0.0]	<u>0.71</u>	0.53	0.76	0.38	0.25	<u>23.95</u>	0.1124
	$\lambda_{1-2}$	[1/2, 1/2, 0.0, 0.0, 0.0]	<b>0.64</b>	<b>0.23</b>	0.34	0.20	0.13	<b>24.08</b>	<b>0.1075</b>
	$\lambda_{1-3}$	[1/3, 1/3, 1/3, 0.0, 0.0]	0.78	0.42	<u>0.10</u>	0.04	0.03	23.81	0.1112
	$\lambda_{1-4}$	[1/4, 1/4, 1/4, 1/4, 0.0]	0.78	0.46	0.18	<b>0.02</b>	<b>0.01</b>	23.68	<u>0.1110</u>
		ESRGAN-OOS $_A$						24.03	0.0848
		ESRGAN-OOS $_B$						24.21	0.0848

Table 3.2: Comparison with state-of-the-art SR methods on benchmarks. The 1st and the 2nd best performances for each group are highlighted in **bold** and underline, respectively. LR-PSNR values greater than 45dB are written in *italic*.

		Distortion-oriented SR				Perception-oriented SR									
Model	Training dataset	SRROE (T = 0)		SRGAN	ESRGAN [?] [?]	SFTGAN [?]	RankSR [?] [?]	SRFlow [?] [?]	SFSR [?] [?]	FxSR [?]	SROOE (T <sub>B</sub> )		SROOS (T <sub>S</sub> )		
		RRDB [?]	DF2K	DF2K	DF2K+OST	ImageNet+OST	DF2K	DF2K	DF2K	DF2K	DF2K	DF2K	DF2K	DF2K	
BSD100	PSNR↑	<b>26.53</b>	<u>26.45</u>	24.13	23.95	24.09	24.09	24.66	24.16	24.77	24.78	<b>24.87</b>	25.07		
	SSIM↑	<b>0.7438</b>	<u>0.7416</u>	0.6454	0.6463	0.6460	0.6438	0.6580	0.6531	0.6817	<u>0.6818</u>	<b>0.6869</b>	0.6960		
	LPIPS↓	<u>0.3575</u>	<b>0.3546</b>	0.1777	0.1615	0.1710	0.1750	0.1833	0.1613	0.1572	<u>0.1530</u>	<b>0.1500</b>	0.1388		
	DISTS↓	<u>0.2005</u>	<b>0.1996</b>	0.1288	0.1158	0.1224	0.1252	0.1372	0.1165	0.1160	<u>0.1139</u>	<b>0.1124</b>	0.1104		
	LR-PSNR↑	<b>52.52</b>	<u>52.35</u>	39.32	41.35	40.92	41.33	<b>40.86</b>	40.99	<i>49.24</i>	<u>48.75</u>	<i>49.19</i>	<i>49.35</i>		
General100	PSNR↑	<b>30.30</b>	<u>30.08</u>	27.54	27.53	27.04	27.31	27.83	27.65	28.44	<u>28.57</u>	<b>28.74</b>	29.12		
	SSIM↑	<b>0.8696</b>	<u>0.8662</u>	0.7998	0.7984	0.7861	0.7899	0.7951	0.7995	0.8229	<u>0.8250</u>	<b>0.8297</b>	0.8400		
	LPIPS↓	<u>0.1665</u>	<b>0.1658</b>	0.0961	0.0880	0.1084	0.0960	0.0962	0.0866	0.0784	<u>0.0764</u>	<b>0.0753</b>	0.0682		
	DISTS↓	<u>0.1321</u>	<b>0.1311</b>	0.0955	0.0845	0.1166	0.0938	0.1022	0.0857	0.0831	<u>0.0811</u>	<b>0.0795</b>	0.0783		
	LR-PSNR↑	<b>53.94</b>	<u>52.79</u>	41.44	41.93	40.05	41.84	49.59	42.30	<i>49.82</i>	<u>49.90</u>	<i>50.11</i>	<i>50.57</i>		
Urban100	PSNR↑	<b>25.48</b>	25.21	22.84	22.78	22.74	22.93	23.68	23.24	24.08	24.21	<b>24.33</b>	24.53		
	SSIM↑	<b>0.8097</b>	<u>0.8020</u>	0.7196	0.7214	0.7107	0.7169	0.7316	0.7365	0.7641	<u>0.7680</u>	<b>0.7707</b>	0.7784		
	LPIPS↓	<u>0.1960</u>	<b>0.1961</b>	0.1426	0.1230	0.1343	0.1385	0.1272	0.1190	0.1090	<u>0.1066</u>	<b>0.1065</b>	0.0988		
	DISTS↓	<u>0.1417</u>	<b>0.1409</b>	0.1001	0.0818	0.0974	0.0987	0.0978	0.0798	0.0783	<u>0.0773</u>	<b>0.0764</b>	0.0774		
	LR-PSNR↑	<b>51.21</b>	<u>50.52</u>	38.84	39.70	39.39	39.07	<b>49.60</b>	40.40	<i>48.27</i>	<u>48.29</u>	<i>48.32</i>	<i>48.52</i>		
Manga109	PSNR↑	<b>29.74</b>	<u>29.36</u>	26.26	26.50	26.07	26.04	27.11	26.74	27.64	27.85	<b>28.08</b>	28.61		
	SSIM↑	<b>0.8997</b>	<u>0.8948</u>	0.8285	0.8245	0.8182	0.8117	0.8244	0.8267	0.8440	<u>0.8493</u>	<b>0.8554</b>	0.8737		
	LPIPS↓	<u>0.0975</u>	<b>0.0972</b>	0.0709	0.0654	0.0716	0.0773	0.0663	0.0683	0.0580	<u>0.0566</u>	<b>0.0524</b>	0.0431		
	DISTS↓	<u>0.0643</u>	<b>0.0605</b>	0.0461	0.0397	0.0496	0.0488	0.0501	0.0403	0.0407	<u>0.0382</u>	<b>0.0351</b>	0.0344		
	LR-PSNR↑	<b>51.73</b>	<u>50.39</u>	40.35	40.68	38.96	39.83	48.36	41.51	48.19	<u>48.49</u>	<b>48.77</b>	49.33		
DIV2K	PSNR↑	<b>29.48</b>	<u>29.33</u>	26.63	26.64	26.56	26.51	27.08	26.71	27.51	<u>27.57</u>	<b>27.69</b>	28.03		
	SSIM↑	<b>0.8444</b>	<u>0.8413</u>	0.7625	0.7640	0.7578	0.7526	0.7558	0.7614	0.7890	<u>0.7906</u>	<b>0.7932</b>	0.8031		
	LPIPS↓	<u>0.2537</u>	<b>0.2530</b>	0.1263	0.1154	0.1449	0.1217	0.1201	0.1100	0.1028	<u>0.0971</u>	<b>0.0957</b>	0.0888		
	DISTS↓	<u>0.1261</u>	<b>0.1254</b>	0.0613	0.0550	0.0858	0.0589	0.0622	0.0494	0.0513	<u>0.0492</u>	<b>0.0491</b>	0.0480		
	LR-PSNR↑	<b>53.71</b>	<u>53.59</u>	40.87	42.61	40.40	41.90	49.96	42.57	<i>50.54</i>	<u>50.36</u>	<b>50.80</b>	<i>51.04</i>		

Table 3.3: Comparison of performance according to different selections. The bold checkmark indicates the change from the left selection. The 1st and the 2nd best performances except for SROOS are highlighted in **bold** and underline, respectively.

Methods		SROT	SROOE			SROOS
Objective	P2	✓	✓			
Trajectory	P1234			✓	✓	✓
<b>T</b>	$\mathbf{T}_{t=0.8}$	✓				
	$\hat{\mathbf{T}}_B$ (OOE)		✓	✓	✓	
	$\mathbf{T}_S^*$ (OOS)					✓
training	DIV2K	✓	✓	✓		
DB	DF2K				✓	✓
Metric	PSNR↑	26.49	26.53	<u>26.65</u>	<b>26.74</b>	27.07
	SSIM↑	0.7803	0.7816	<u>0.7853</u>	<b>0.7872</b>	0.7982
	LPIPS↓	0.1011	0.1000	<u>0.0978</u>	<b>0.0960</b>	0.0875
	LR-PSNR↑	49.21	49.21	<u>49.32</u>	<b>49.44</b>	49.76

Table 3.4: Comparison of the running time and the SR model size.

	SRGAN [?]	ESRGAN [?]	FxSR [?]	SROOE
Run Time (msec)	0.014	0.138	0.501	0.968
Param Size (MB)	1.51	16.69	18.30	70.20

## Chapter 4

### Conclusions

This dissertation presents a perception-oriented image restoration method using conditional objective learning. It is shown that the results of the proposed method overcome the trade-off in image restoration performance by adaptively applying distortion-oriented and perceptual-oriented results to each region. Specifically, two key methods are proposed: an efficient method to train a single locally-adjustable model using conditional objective, and a novel IR framework that estimates and applies an optimal combination of objectives for each input region and thus produces perceptually accurate restoration results.

First, We have presented a novel training method and a network structure for the SISR, enabling us to explore various region-wise HR outputs. From this, we can flexibly reconstruct the images between perception-oriented and distortion-oriented ones. This is achieved by defining a conditional objective function with the weights related to the perceptual losses in various feature space levels. Also, our network is designed to modulate the network's intermediate features to change the operation according to these control inputs. As a result, we can generate an image with a desired restoration style for each area. Experiments show that the proposed FxSR yields state-of-the-art perceptual quality and higher PSNR than other perception-oriented methods. Also, we can find many solutions by controlling a single parameter at the inference phase. We

will release our code for further research and comparisons.

Second, this dissertation has proposed a novel image restoration framework for perceptually accurate HR restoration, where an objective estimator provides an optimal combination of objectives for a given image patch, and a generator produces SR results that reflect the target objectives. For this purpose, we employed objective trajectory learning to efficiently train a single generative model that can apply varying objective sets. Experiments show that the proposed method reduces visual artifacts, such as structural distortion and unnatural details, and achieves improved results compared to those of state-of-the-art perception-oriented methods in terms of both perceptual and distortion metrics. The proposed method can be applied to off-the-shelf and other SISR network architectures.



# Bibliography

- [1] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, “Esrgan: Enhanced super-resolution generative adversarial networks,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0. 2, 9, 10, 13, 14, 22, 23, 24, 25, 26, 27, 29, 32, 42, 43, 45, 48, 57, 58, 61, 62
- [2] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, “Enhancenet: Single image super-resolution through automated texture synthesis,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4491–4500. 2, 13, 43, 47
- [3] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, “Learning to super-resolve blurry face and text images,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 251–260. 2, 9, 43
- [4] Y. Blau and T. Michaeli, “The perception-distortion tradeoff,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6228–6237. 2, 25, 26, 46
- [5] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, “Deep learning for single image super-resolution: A brief review,” *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019. 6

- [6] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in *European conference on computer vision*. Springer, 2014, pp. 184–199. 6, 41, 45
- [7] C. Dong, C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015. 6, 41
- [8] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1646–1654. 6, 45
- [9] ———, “Deeply-recursive convolutional network for image super-resolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645. 6
- [10] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883. 6
- [11] Y. Tai, J. Yang, X. Liu, and C. Xu, “Memnet: A persistent memory network for image restoration,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4539–4547. 6, 41
- [12] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144. 6, 41, 45, 57
- [13] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in *Proceedings of the Eu-*

- ropean Conference on Computer Vision (ECCV), 2018, pp. 286–301. 6, 41, 45
- [14] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481. 6, 41, 45
- [15] X. Yang, H. Mei, J. Zhang, K. Xu, B. Yin, Q. Zhang, and X. Wei, “Drfn: Deep recurrent fusion network for single-image super-resolution with large factors,” *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 328–337, 2018. 6
- [16] Z. Jin, M. Z. Iqbal, D. Bobkov, W. Zou, X. Li, and E. Steinbach, “A flexible deep cnn framework for image restoration,” *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1055–1068, 2019. 6
- [17] Y. Zhang, P. Wang, F. Bao, X. Yao, C. Zhang, and H. Lin, “A single-image super-resolution method based on progressive-iterative approximation,” *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1407–1422, 2019. 6
- [18] Z. He, Y. Cao, L. Du, B. Xu, J. Yang, Y. Cao, S. Tang, and Y. Zhuang, “Mrfn: Multi-receptive-field network for fast and accurate single image super-resolution,” *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1042–1054, 2019. 6
- [19] X. Zhang, P. Gao, S. Liu, K. Zhao, G. Li, L. Yin, and C. W. Chen, “Accurate and efficient image super-resolution via global-local adjusting dense network,” *IEEE Transactions on Multimedia*, 2020. 6
- [20] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C.-W. Lin, “Coarse-to-fine cnn for image super-resolution,” *IEEE Transactions on Multimedia*, vol. 23, pp. 1489–1502, 2020. 6

- [21] M. Mathieu, C. Couprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” in *4th International Conference on Learning Representations, ICLR 2016 ; Conference date: 02-05-2016 Through 04-05-2016*, Jan. 2016. 6
- [22] A. Dosovitskiy and T. Brox, “Generating images with perceptual similarity metrics based on deep networks,” in *Advances in neural information processing systems*, 2016, pp. 658–666. 6, 9
- [23] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European conference on computer vision*. Springer, 2016, pp. 694–711. 6, 9, 41, 45
- [24] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690. 6, 9, 13, 22, 25, 31, 32, 39, 41, 42, 45, 47, 57, 61, 62
- [25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680. 6, 8, 16, 41, 45
- [26] X. Wang, K. Yu, C. Dong, and C. Change Loy, “Recovering realistic texture in image super-resolution by deep spatial feature transform,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 606–615. 7, 9, 13, 14, 15, 22, 24, 25, 43, 44, 45, 54, 57, 61
- [27] M. S. Rad, B. Bozorgtabar, U.-V. Marti, M. Basler, H. K. Ekenel, and J.-P. Thiran, “Srobb: Targeted perceptual loss for single image super-resolution,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2710–2719. 7, 9, 13, 43, 47

- [28] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, “SrfLOW: Learning the super-resolution space with normalizing flow,” in *European Conference on Computer Vision*. Springer, 2020, pp. 715–732. 7, 22, 25, 28, 29, 57, 61
- [29] W. Wang, R. Guo, Y. Tian, and W. Yang, “Cfsnet: Toward a controllable feature space for image restoration,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4140–4149. 7, 10
- [30] X. Wang, K. Yu, C. Dong, X. Tang, and C. C. Loy, “Deep network interpolation for continuous imagery effect transition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1692–1701. 7, 10, 28, 43
- [31] A. Shoshan, R. Mechrez, and L. Zelnik-Manor, “Dynamic-net: Tuning the objective without re-training for synthesis tasks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3215–3223. 7, 10
- [32] E. L. Denton, S. Chintala, R. Fergus *et al.*, “Deep generative image models using a laplacian pyramid of adversarial networks,” in *Advances in neural information processing systems*, 2015, pp. 1486–1494. 9
- [33] H. Ren, A. Kheradmand, M. El-Khamy, S. Wang, D. Bai, and J. Lee, “Real-world super-resolution using generative adversarial networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 436–437. 9
- [34] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 126–135. ix, x, 9, 18, 19, 24, 25, 26, 48, 57
- [35] J. Bruna, P. Sprechmann, and Y. LeCun, “Super-resolution with deep convolutional sufficient statistics,” in *4th International Conference on Learning Rep-*

- resentations, *ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. 9
- [36] Z. Hui, J. Li, X. Gao, and X. Wang, “Progressive perception-oriented network for single image super-resolution,” *Information Sciences*, vol. 546, pp. 769–786, 2021. 9
- [37] K. Zhang, S. Gu, and R. 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 Workshops*, 2020, pp. 492–493. 9, 31
- [38] T. Tariq, O. T. Tursun, M. Kim, and P. Didyk, “Why are deep representations good perceptual quality features?” in *European Conference on Computer Vision*. Springer, 2020, pp. 445–461. 9
- [39] A. Lugmayr, M. Danelljan, R. Timofte, M. Fritsche, S. Gu, K. Purohit, P. Kandula, M. Suin, A. Rajagoapalan, N. H. Joon *et al.*, “Aim 2019 challenge on real-world image super-resolution: Methods and results,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3575–3583. 9
- [40] Y. Jo, S. Yang, and S. Joo Kim, “Investigating loss functions for extreme super-resolution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 424–425. 9
- [41] B. Yan, B. Bare, C. Ma, K. Li, and W. Tan, “Deep objective quality assessment driven single image super-resolution,” *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2957–2971, 2019. 9
- [42] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in *Proceedings of the 32nd In-*

- ternational Conference on Machine Learning*, ser. JMLR Proceedings, vol. 37. JMLR.org, 2015, pp. 448–456. 9, 41
- [43] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, “Instance normalization: The missing ingredient for fast stylization,” *CoRR*, vol. abs/1607.08022, 2016. 9
- [44] V. Dumoulin, J. Shlens, and M. Kudlur, “A learned representation for artistic style,” *CoRR*, vol. abs/1610.07629, 2016. 9
- [45] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510. 9
- [46] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 9, 11
- [47] J. He, C. Dong, and Y. Qiao, “Modulating image restoration with continual levels via adaptive feature modification layers,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 056–11 064. 10
- [48] Y. Bahat and T. Michaeli, “Explorable super resolution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2716–2725. 10
- [49] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997. 10, 11
- [50] J. Baxter, “A model of inductive bias learning,” *Journal of artificial intelligence research*, vol. 12, pp. 149–198, 2000. 11
- [51] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *IEEE Transactions on Knowledge and Data Engineering*, 2021. 11

- [52] A. Dosovitskiy and J. Djolonga, “You only train once: Loss-conditional training of deep networks,” in *International Conference on Learning Representations*, 2019. 11, 47, 52
- [53] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. 11, 15, 47, 56
- [54] A. Jolicoeur-Martineau, “The relativistic discriminator: a key element missing from standard GAN,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 16
- [55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. ix, 19, 20, 23, 25, 45, 57
- [56] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402. ix, 19, 25
- [57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595. ix, xi, 20, 21, 23, 25, 42, 43, 50, 57
- [58] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020. ix, 20, 23, 25, 57



- [59] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006. ix, 20, 25, 27
- [60] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ”completely blind” image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013. ix, 20, 23, 25
- [61] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012. ix, 20, 25
- [62] A. Lugmayr, M. Danelljan, and R. Timofte, “Unsupervised learning for real-world super-resolution,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3408–3416. 20
- [63] Learning the super-resolution space challenge, ntire 2021 at cvpr. [Online]. Available: [https://github.com/andreas128/NTIRE21\\_Learning\\_SR\\_Space](https://github.com/andreas128/NTIRE21_Learning_SR_Space) 21, 29
- [64] A. Lugmayr, M. Danelljan, and R. Timofte, “Ntire 2021 learning the super-resolution space challenge,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 596–612. xi, 21, 29, 42, 50, 57
- [65] J. W. Soh, G. Y. Park, J. Jo, and N. I. Cho, “Natural and realistic single image super-resolution with explicit natural manifold discrimination,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8122–8131. 22, 24, 25, 43, 45
- [66] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, “Structure-preserving super resolution with gradient guidance,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7769–7778. 22, 24, 25, 42, 43, 46, 57, 61

- [67] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1412.6980> 23, 41
- [68] A. Saxena, S. H. Chung, A. Y. Ng *et al.*, “Learning depth from single monocular images,” in *NIPS*, vol. 18, 2005, pp. 1–8. x, 30, 38
- [69] A. Saxena, S. H. Chung, and A. Y. Ng, “3-d depth reconstruction from a single still image,” *International journal of computer vision*, vol. 76, no. 1, pp. 53–69, 2008. 30
- [70] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, “Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera,” in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 559–568. 30
- [71] Y. Cui, S. Schuon, S. Thrun, D. Stricker, and C. Theobalt, “Algorithms for 3d shape scanning with a depth camera,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 5, pp. 1039–1050, 2012. 30
- [72] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt, “3d shape scanning with a time-of-flight camera,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1173–1180. 30
- [73] Y. Ming, X. Meng, C. Fan, and H. Yu, “Deep learning for monocular depth estimation: A review,” *Neurocomputing*, vol. 438, pp. 14–33, 2021. 30
- [74] A. Lugmayr, M. Danelljan, and R. 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 Workshops*, 2020, pp. 494–495. 31
- [75] S. Nah, R. Timofte, S. Gu, S. Baik, S. Hong, G. Moon, S. Son, and K. Mu Lee, “Ntire 2019 challenge on video super-resolution: Methods and results,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0. 31
- [76] S. A. Hussein, T. Tirer, and R. Giryes, “Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1428–1437. 31
- [77] N. Ahn, J. Yoo, and K.-A. Sohn, “Simusr: A simple but strong baseline for unsupervised image super-resolution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 474–475. 31
- [78] 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)*. IEEE, 2019, pp. 3599–3608. 31
- [79] K. Zhang, W. Zuo, and L. Zhang, “Learning a single convolutional super-resolution network for multiple degradations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3262–3271. 31
- [80] A. Lugmayr, M. Danelljan, and R. Timofte, “Unsupervised learning for real-world super-resolution,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3408–3416. 31
- [81] K. Deb, “Multi-objective optimisation using evolutionary algorithms: an introduction,” in *Multi-objective evolutionary optimisation for product design and manufacturing*. Springer, 2011, pp. 3–34. 32

- [82] X. Lin, H.-L. Zhen, Z. Li, Q.-F. Zhang, and S. Kwong, “Pareto multi-task learning,” *Advances in neural information processing systems*, vol. 32, pp. 12 060–12 070, 2019. 32
- [83] S. C. Park, M. K. Park, and M. G. Kang, “Super-resolution image reconstruction: a technical overview,” *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003. 41
- [84] H. Greenspan, “Super-resolution in medical imaging,” *The Computer Journal*, vol. 52, no. 1, pp. 43–63, 2009. 41
- [85] C. You, G. Li, Y. Zhang, X. Zhang, H. Shan, M. Li, S. Ju, Z. Zhao, Z. Zhang, W. Cong, M. W. Vannier, P. K. Saha, E. A. Hoffman, and G. Wang, “Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle),” *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 188–203, 2020. 41
- [86] T. Uiboupin, P. Rasti, G. Anbarjafari, and H. Demirel, “Facial image super resolution using sparse representation for improving face recognition in surveillance monitoring,” in *2016 24th Signal Processing and Communication Application Conference (SIU)*, 2016, pp. 437–440. 41
- [87] M. Farooq, M. N. Dailey, A. Mahmood, J. Moonrinta, and M. Ekpanyapong, “Human face super-resolution on poor quality surveillance video footage,” *Neural Computing and Applications*, vol. 33, no. 20, pp. 13 505–13 523, 2021. 41
- [88] Y. Luo, L. Zhou, S. Wang, and Z. Wang, “Video satellite imagery super resolution via convolutional neural networks,” *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2398–2402, 2017. 41
- [89] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, “Spatiotemporal satellite image fusion using deep convolutional neural networks,” *IEEE Journal of Se-*

*lected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 821–829, 2018. 41

- [90] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep laplacian pyramid networks for fast and accurate super-resolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632. 41
- [91] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *European conference on computer vision*. Springer, 2016, pp. 630–645. 41
- [92] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, “Ranksrgan: Generative adversarial networks with ranker for image super-resolution,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3096–3105. 42, 43, 46, 57, 61
- [93] S. H. Park, Y. S. Moon, and N. I. Cho, “Flexible style image super-resolution using conditional objective,” *IEEE Access*, vol. 10, pp. 9774–9792, 2022. xi, 43, 47, 50, 51, 52, 54, 57, 58, 59, 61, 62
- [94] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. 45
- [95] N. Ahn, B. Kang, and K.-A. Sohn, “Fast, accurate, and lightweight super-resolution with cascading residual network,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 45
- [96] M. Haris, G. Shakhnarovich, and N. Ukita, “Deep back-project networks for single image super-resolution,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4323–4337, 2021. 45

- [97] S. Anwar and N. Barnes, “Densely residual laplacian super-resolution,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1192–1204, 2022. 45
- [98] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Swinir: Image restoration using swin transformer,” in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 1833–1844. 45
- [99] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, “Uformer: A general u-shaped transformer for image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 17 683–17 693. 45
- [100] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 012–10 022. 45
- [101] Z. Wang, J. Chen, and S. C. H. Hoi, “Deep learning for image super-resolution: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2021. 45
- [102] J. Liang, H. Zeng, and L. Zhang, “Details or artifacts: A locally discriminative learning approach to realistic image super-resolution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5657–5666. 46
- [103] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, “The 2018 pfirm challenge on perceptual image super-resolution,” in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018. 46

- [104] 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 *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001, pp. 416–423 vol.2. 48, 57
- [105] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in *European conference on computer vision*. Springer, 2016, pp. 391–407. 48, 57
- [106] J.-B. Huang, A. Singh, and N. Ahuja, “Single image super-resolution from transformed self-exemplars,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5197–5206. 48, 49, 57
- [107] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, “Sketch-based manga retrieval using manga109 dataset,” *Multimedia Tools and Applications*, vol. 76, no. 20, pp. 21 811–21 838, 2017. 48, 57
- [108] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241. 56

# 초 록

영상 복원의 목적은 주어진 저품질 영상을 고품질 영상으로 복원하는 것이다. 전형적인 영상 복원 분야에는 영상 잡음 제거(image denoising)와 영상 초해상화(image super-resolution)가 포함된다. 영상 복원은 일상의 영상 품질 향상뿐 아니라 의료, 감시 및 위성 이미지의 컴퓨터 비전 작업 전처리 단계로도 많이 활용되고 있다. 그러나 이러한 영상 복원은 하나의 저품질 이미지에 무한히 많은 고품질 이미지들이 대응한다는 점에서 불량조건문제이기 때문에 어려운 작업이다. 최근에는 대규모 외부 데이터 세트로 훈련된 심층 신경망을 도입하여 영상 복원의 성능이 크게 향상되었다. 특히 픽셀 단위의 왜곡 감소 지향 손실(L1 및 L2)은 높은 신호 대 잡음비(PSNR)를 얻는데 도움이 되며 초기 연구부터 널리 사용되었다. 그러나 이러한 왜곡 기반 손실로 학습된 모델은 주어진 저품질 이미지에 대응될 수 있는 고품질 솔루션들의 평균을 복원 결과로 생성하게 되며, 이는 일반적으로 흐릿하고 시각적으로 만족스럽지 않다. 이후 이러한 문제를 극복하고 세밀한 디테일이 있는 실제같은 이미지를 생성하기 위해 인지 손실(perceptual loss) 및 생성 적대적 손실(generative adversarial loss)과 같은 인지 지향 손실들이 도입되었다. 이러한 인지 지향 손실은 다양한 영상 복원 방법들에 사용되었지만, 부자연스러운 디테일 및 구조적 왜곡의 발생과 같은 바람직하지 않은 부작용도 가져온다. 특히 단일 인지 손실을 영상 전체에 동일하게 사용하는 것은 국부적으로 다양한 형태를 가지는 이미지를 정확하게 복원하는데 충분하지 않은 것으로 나타났다. 이러한 이유로 인지 손실, 적대적 손실, 왜곡 손실 등 다양한 손실들의 가중 조합이 시도되었지만 최적의 조합을 찾는 것은 여전히 어려운 일이다. 이러한 문제를 해결하기 위해 본 학위 논문에서는 지역별로 최적 목적을 예측 및 적용하여 복원 결과 영상의 전체 영역에서 실제같고 자연스러운 결과를 생



성하는 새로운 방법을 제시한다.

첫 번째 연구는 제어 맵에 따라 국부적으로 다양한 형태의 고해상도 복원 결과를 생성할 수 있는 유연한 모델의 학습법과 네트워크 구조이다. 일반적으로, 다양한 초해상화 결과를 얻기 위한 접근 방식은 손실 가중치가 다른 여러 목적들로 각각의 모델을 훈련하고 이러한 모델들의 조합을 활용하는 것이지만, 본 연구에서는 여러 모델을 사용하는 대신, 훈련 중에 조건 목적으로 단일의 초해상화 모델을 최적화하는 방법을 제안한다. 여기서 목적은 각각 다른 비전 레벨의 특징에 해당되는 인지 손실 항들의 가중 합을 포함한다. 이 가중치 집합은 스타일 제어 입력에 따라 다르게 정의된다. 또한, 이 훈련 방식에 적합한 네트워크 구조로 공간 특징 변환 레이어가 장착된 Residual-in-Residual Dense Block을 제시한다. 이렇게 훈련된 모델은, 추론 단계에서, 국부적으로 변화하는 목적 맵에 대응되는 고해상도 복원 결과를 생성할 수 있다. 광범위한 실험은 제안된 초해상화 모델이 부작용 없이 목적 제어 맵에 따라 다양한 스타일의 초해상화 복원 결과를 생성하고 최첨단 초해상화 방법들에 필적하는 정량적 성능도 달성한다는 것을 보여준다.

두 번째 연구에서는, 인지 관점에서 최적의 목적을 지역마다 추정하고 이를 적용함으로써 복원 영상 전체 영역에서 고품질을 달성할 수 있는, 인지 지향의 새로운 영상 복원 프레임워크를 제시한다. 구체적으로 프레임워크는 주어진 저해상도 입력에 대한 최적의 목적 맵을 유추하는 예측 모델과 해당 목적 맵에 상응하는 초해상화 복원 결과를 생성하는 생성 모델의 두 가지 모델로 구성된다. 생성 모델은 제안하는 필수 목적들을 포함하는 목적 궤적에 대해 훈련되며, 이를 통해 단일 생성 모델은 연속된 궤적 상의 다양하게 결합된 손실들에 해당하는 다양한 초해상화 결과들을 학습할 수 있다. 예측 모델은 저해상도 이미지와 그에 상응하는 목적 궤적에서 검색된 최적의 목적 맵의 쌍을 사용하여 훈련된다. 5개의 벤치마크에 대한 실험 결과는 제안하는 방법이 LPIPS, DISTs, PSNR 및 SSIM 측정에서 최신 인지 기반 초해상화 방법들보다 성능이 우수함을 보여준다. 또한 시각적 비교 결과에서도 인지 지향 복원 관점에서 제안 방법의 우수성을 보여준다.

**주요어:** 영상 복원, 이미지 초해상화, 인지 지향 영상 복원, 인지 지향 이미지 초해상화, 조건 목적, 최적 목적 추정

**학번:** 2015-31015