Structure-aware Meta-fusion for Image Super-resolution

HAOYU MA, BINGCHEN GONG, and YIZHOU YU, The University of Hong Kong, China

There are two main categories of image super-resolution algorithms: distortion oriented and perception oriented. Recent evidence shows that reconstruction accuracy and perceptual quality are typically in disagreement with each other. In this article, we present a new image super-resolution framework that is capable of striking a balance between distortion and perception. The core of our framework is a deep fusion network capable of generating a final high-resolution image by fusing a pair of deterministic and stochastic images using spatially varying weights. To make a single fusion model produce images with varying degrees of stochasticity, we further incorporate meta-learning into our fusion network. Once equipped with the kernel produced by a kernel prediction module, our meta fusion network is able to produce final images at any desired level of stochasticity. Experimental results indicate that our meta fusion network outperforms existing state-of-the-art SISR algorithms on widely used datasets, including PIRM-val, DIV2K-val, Set5, Set14, Urban100, Manga109, and B100. In addition, it is capable of producing high-resolution images that achieve low distortion and high perceptual quality simultaneously.

CCS Concepts: • **Computing methodologies** → **Image processing**; *Machine learning approaches;*

Additional Key Words and Phrases: Super-resolution, meta-learning, image fusion

ACM Reference format:
Haoyu Ma, Bingchen Gong, and Yizhou Yu. 2022. Structure-aware Meta-fusion for Image Super-resolution. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 2, Article 60 (February 2022), 25 pages.
https://doi.org/10.1145/3477553

1 INTRODUCTION

Single image super-resolution (SISR) recovers **high-resolution (HR)** images from single low-resolution ones. As a fundamental low-level computer vision problem, it has been popular in the past decades and intrigued vast research interests. In recent years, the introduction of deep learning also provides a significant performance boost to super-resolution models. Due to the convenience in transferring deep learning models, tailored super-resolution algorithms have been applied to many domain-specific problems to boost the performance of existing algorithms, such as tiny object detection [3], video super-resolution/enhancement [14, 18], remote sensing [47], and blind image denoising [13].

The super-resolution problem is ill posed in essence: One low-resolution image has many corresponding high-resolution images at the same time, because part of the high-frequency

This work was supported in part by Hong Kong PhD Fellowship and Hong Kong Research Grants Council through Research Impact Fund under Grant R-5001-18.
Haoyu Ma and Bingchen Gong contributed equally to this research.
Authors' address: H. Ma, B. Gong, and Y. Yu, The University of Hong Kong, China; emails: mahaoyu@connect.hku.hk, bccs@connect.hku.hk, yizhouy@acm.org.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
1551-6857/2022/02-ART60 $15.00
https://doi.org/10.1145/3477553

information is forever lost due to the low sampling rate of the low-resolution image. Therefore, a high-resolution image can be decomposed into a deterministic component and a stochastic component. The deterministic component refers to the information in the image highly correlated with the low-resolution image and can be recovered by super-resolution algorithms. It is strongly related to salient structures and smooth regions. Besides those salient structures, the perceptual quality of super-resolved images is closely related to the high-frequency information in regions with textures and patterns. Part of such high-frequency information in the original high-resolution image cannot be recovered due to downsampling, and the lost high-frequency information forms the stochastic component, making image super-resolution an ill-posed problem.

Image super-resolution algorithms generally fall into two main categories: distortion oriented and perception oriented. Distortion-oriented algorithms strive to lower reconstruction errors and primarily focus on recovering deterministic components, as stochastic components are considered as noises during reconstruction and often increase reconstruction errors. However, stochastic details tend to increase the perceived resolution of an image. Perception-oriented algorithms aim to synthesize vivid stochastic components to make super-resolution images look natural and detailed. Nevertheless, synthesized stochastic details sometimes significantly deviate from the ground truth.

Evidences in recent advancements [6] show that reconstruction accuracy and perceptual quality are typically in disagreement with each other. Distortion-oriented models tend to produce perceptually unpleasant noise-free results while perception-oriented algorithms usually undermine distortion metrics, such as **Peak Signal Noise Ratio(PSNR)/Root Mean Square Error(RMSE)**. Although efforts have been made to push the boundary of these two main categories, there is a clear gap between distortion-oriented and perception-oriented models. This disagreement has become a fundamental tradeoff between distortion and natural image statistics. How to introduce stochastic details while still maintaining low distortion remains a significant challenge.

Although stochastic details are necessary to make a super-resolved image perceptually better, the magnitude of the stochastic component to be introduced into an image depends on the actual content as well as the user of the content. Some types of images, such as images of natural scenes with grass and trees, can tolerate a large stochastic component while the others, such as images of indoor scenes filled with human-made objects, may prefer a much smaller one. For the same reason, different users of the super-resolution content may prefer different degrees of stochasticity. Since it is cumbersome to train many super-resolution models of different degrees of stochasticity, a single model capable of introducing varying degrees of stochasticity under the control of an input signal is desired.

Varying degrees of stochasticity are also wanted for pixels in the same image. For instance, it is necessary to distinguish pixels on salient structures, such as edges and contours, from the remaining ones. Salient structures exhibit clear and spatially coherent orientations while the orientation at other pixels, including those in texture regions, is more ambiguous or spatially incoherent. This distinction implies that the deterministic component plays a much more dominant role than the stochastic component around salient structures while the stochastic component becomes more important at other pixels. We call pixels on salient structures structural pixels and the rest non-structural pixels.

In this article, we present a new image super-resolution framework that is capable of striking an excellent balance between distortion and perception. Our framework leverages meta-learning as well as existing state-of-the-art SISR algorithms. Its core is a deep fusion network capable of performing spatially varying image fusion. It takes as input two high-resolution images representing the deterministic and stochastic components of an underlying ground-truth high-resolution image. These two input images are respectively generated by distortion-oriented and perception-oriented SISR algorithms. The fusion network then generates a final high-resolution image by fusing the

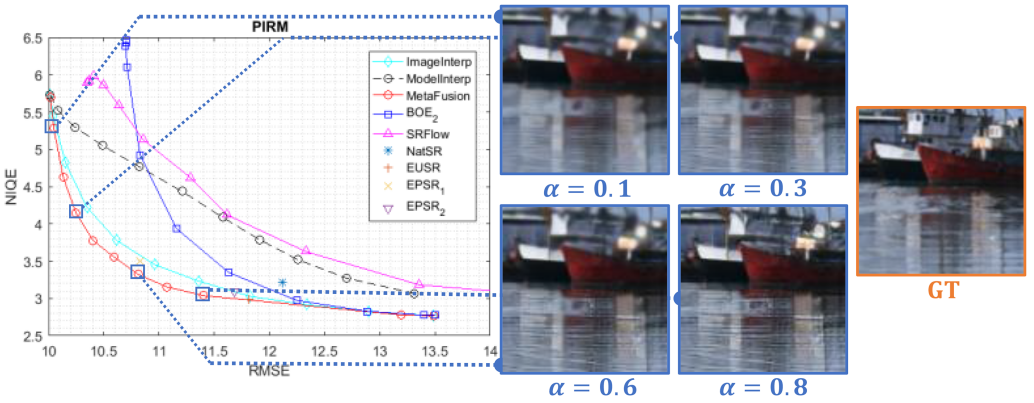

Fig. 1. Left: Perception–distortion plane for PIRM self-validation dataset. We also show location of other models such as NatSR [30], BOE [26], EUSR [8], EPSR [36] image interpolated results from ESRGAN[42] and model interplolated results from ModelInterp [41]. Right: By preserving structure regions and controlling the stochasticity level, we produce visually satisfactory images with lower distortion.

two input images using spatially varying weights for structural and non-structural pixels. To make the fusion network aware of the spatial distribution of structural and non-structural pixels, a structure mapping module is designed to produce a structure map, which is concatenated with the two input images before fusion is performed.

To make a single fusion model capable of producing images with varying degrees of stochasticity, we further incorporate meta-learning into our fusion network. This is achieved by transforming one of the convolutional layers in the fusion network into a meta convolutional layer, whose kernel is not trained but predicted by a separate kernel prediction module. The input to this kernel prediction module is a single parameter called stochastic weight, which controls the desired level of stochasticity in the final high-resolution image. Once equipped with the kernel produced by the kernel prediction module, our meta fusion network is able to produce final images at the level of stochasticity defined by the stochastic weight. To train our meta fusion network, we propose a structure-modulated loss, which assembles high-resolution images at varying degrees of stochasticity on the fly using the aforementioned two input images and a structure guide for structural pixels in the ground-truth high-resolution image.

Our proposed model has been tested on widely used datasets, including PIRM-val, DIV2K-val, Set5, Set14, Urban100, Manga109, and B100. Experimental results indicate that our meta fusion network outperforms existing state-of-the-art SISR algorithms and is capable of producing high-resolution images that achieve low distortion and high perceptual quality simultaneously. Furthermore, it is capable of producing images with varying degrees of stochasticity according to the input signal to the kernel prediction module as shown in Figure 1.

In summary, this article has the following three contributions.

- **An efficient fusion network** to integrate input images representing deterministic and stochastic components and produce final high-resolution images that balance distortion and perceptual quality.
- **A meta-learning module** to generate the fusion network on the fly, which can produce images at varying degrees of stochasticity.
- **A structure-modulated loss** to incorporate stochastic components without destroying local deterministic structures during training.

2 RELATED WORK

Earlier exemplar-based methods [7, 11] require a database of external images and synthesize a high-resolution version of the input image by searching for exemplars in the database and transferring relevant patches from the retrieved exemplars. Later dictionary-based methods [34, 35, 46, 48] learn a compact dictionary in feature space to reduce the computational cost of patch-level exemplar retrieval. While the conventional non-DNN methods explicitly utilise external image and feature priors such as edge, gradient [31, 32], and internal non-local similarity [16], the deep learning–based methods implicitly adopt image and feature priors that are inscribed in training datasets, and achieve great breakthrough in accuracy.

Distortion-oriented model. In deep learning, the common approach to solving the super-resolution problem is to design a model capable of generating estimated super-resolved images with the lowest expectation of distortion. SRCNN [9] is a deep network with three convolutional layers designed for super-resolution regression given pairs of low-resolution and high-resolution images. SRResNet [17] and EDSR [19] suggest that multiple deep and wide residual networks can further improve the performance. At present, EDSR acts as a baseline for cutting-edge distortion-oriented models. With the idea of shifting the limited computational resources based on the importance of informative components, RCAN [50] incorporates channel attention mechanism into a residual in residual structured network. SAN outperforms the RCAN by exploiting the inherent feature correlations in intermediate layers with the second-order attention network. SCAN [45], compared against RCAN, developes a new colour attention mechanism and shows performance improvements on real SISR competition dataset. Another state-of-the-art architecture in Reference [20] is also invented to minimize the distortion by aggregating the features and focusing more on the critical contents. Other frameworks are also invented to tackle downstream distortion-oriented tasks such as blind super-resolution [38, 49], arbitrary scale super-resolution [39], and fast inference [37].

Perception-oriented model. The milestone in this area is SRGAN[17], next enhanced by ESRGAN [42], where a stand-alone discriminator is designed to judge whether the image is distinguishable from the real image and penalize on the generator if the image is over-smoothed. Perceptual loss is also used to evaluate the high-level similarity between generated super-resolution images and original high-resolution images to encourage semantic similarities instead of enforcing pixel-wise concurrence. SFT-GAN [40] transforms spatial features to include high-level categorical prior into SRGAN, while NatSR [30] uses low-level domain prior, showing that auxiliary image statistics can be used to improve perceptual quality. Advancements in image stylization also have been adopted in SRNTT [51], where a high-resolution reference image is used to compensate for missing details during super-resolution.

Meta-learning. Known as "learning to learn," meta-learning is a research field that intends to gain experience from a wide range of machine learning tasks. The learned metadata can be adapted or generalized to new tasks or new environments with minimal effort. Some meta-learning applications include few-shot/zero-shot learning [2, 29] and transfer learning [43]. Meta-SR [15] first adopts weight prediction, one of the meta-learning strategies, to perform arbitrary scale super-resolution with one model. In Meta-SR, different kernel weights are calculated given different upscale and pixel locations, achieving better performance on various scale factors compared to baseline. This suggests that it is possible to use meta-learning in a super-resolution generative network and inspires our meta-fusion model design.

The tradeoff between perception and distortion has been considered explicitly in recent works. EPSR [36] controls the tradeoff by carefully balancing weight between reconstruction loss and perception loss. Each tradeoff point requires a delicately chosen weight during training model.

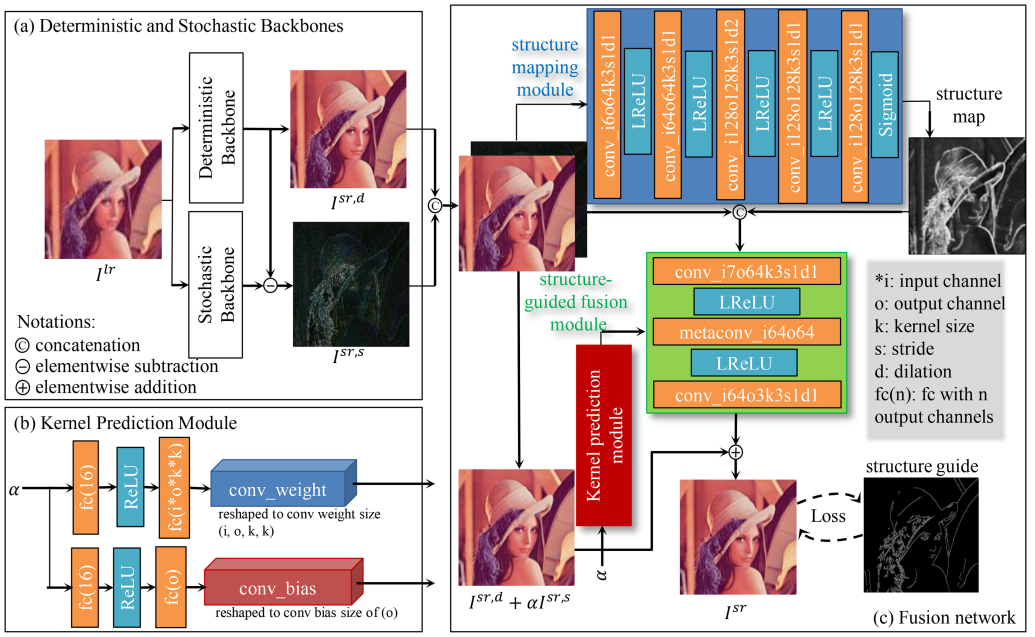

Fig. 2. Pipeline for the proposed meta-fusion algorithm. (a) Deterministic and stochastic backbones. (b) A scalar stochastic weight α is used by kernel prediction module to predict kernel parameters. (c) Deterministic/stochastic images are used by fusion network to generate final output images.

Michelini et al. [26], using an iterative Back-Projections network, can achieve the tradeoff by adjusting the amplitude of input noise. The most relevant work to ours is DTSN [12], where deterministic and stochastic components are generated separately and merged by a fixed network. In this article, the fusion network is generated based on input parameter on the fly. We use a single model to generate HR image in different perception and distortion tradeoff.

3 ALGORITHM

3.1 Overview

The pipeline for our proposed method is shown in Figure 2. Given a low-resolution image I^{lr}, a SISR algorithm reconstructs an estimated high-resolution image I^{sr}. It should be faithful to the ground-truth high-resolution image I^{hr} and as visually pleasant as possible. To achieve both goals, our proposed method has two stages. In the first stage, two HR images, $I^{sr,d}$ and $I^{sr,s}$, are constructed by two super-resolution models trained with different loss functions to estimate a deterministic component and a stochastic component, respectively. The former aims to recover the major structural information of the unknown high-resolution image. The latter aims to hallucinate the high-frequency information lost during down-sampling.

In the second stage, the deterministic and stochastic components are fused together to produce the final high-resolution image. Traditionally, it is done by image interpolation $I^{sr,d} + \alpha I^{sr,s}$. Here α is a scalar that serves as a weight of the stochastic component introduced into the final super-resolution image. However, since α is fixed spatially, in regions dominated by the deterministic component, it inevitably introduces extra noise from the stochastic component and pollute the synthesized high-resolution image.

Therefore, in the second stage, deterministic and stochastic components, $I^{sr,d}$ and $I^{sr,s}$, along with stochastic weight α are taken by our fusion network as input. In particular, α is the input to the kernel prediction module, which is the meta-learning module that generates kernel weight and bias for one of the convolutional layers in our network. As identifying spatially varying structural information plays an important role in spatially varying image fusion, a structure mapping module is designed to predict a structure map of the image. Then the predicted structure map is concatenated with $I^{sr,d}$ and $I^{sr,s}$, and the final super-resolved image is synthesized as $f(I^{sr,s}, I^{sr,d}, \theta_f(\alpha))$, where θ_f represents all network parameters generated by the kernel prediction module. To speed up the convergence of the training stage, we also adopt a residual learning architecture for our fusion network, which is trained by optimizing the structure-modulated loss function described in Section 3.5 in an end-to-end manner.

3.2 Deterministic and Stochastic Component Generation

Although the choices of networks for deterministic and stochastic component generation are not limited, it is expected that the network for deterministic component reconstruction is able to reconstruct images as faithfully as possible and achieve reasonably high PSNRs. Moreover, the network for stochastic component synthesis should be able to hallucinate high-frequency information lost during down-sampling. These two networks serve as the backbones and have a fundamental impact on the final results produced by our fusion network.

In our experiments, we use existing super-resolution networks as our networks for deterministic and stochastic component generation. To facilitate a comparison with interpolation strategies such as naive image interpolation strategy and other state-of-the-art strategies such as model interpolation [41], we use ESRGAN [42] for both deterministic and stochastic component generation. We also use the BOE model [26] to compare the influence of different backbone choices. To ensure pixel-level faithfulness of the deterministic branch, L1 loss is used and model parameters θ_d are trained as $\arg\min_{\theta_d} ||I^{sr,d} - I^{hr}||$. For the stochastic branch, we adopt the original perceptual loss settings described in their works. The stochastic component is calculated by simple subtraction of $I^{sr,d}$ from image generated by the stochastic backbone as shown in Figure 2.

In comparison to traditional one-stage super-resolution models, our design divides the super-resolution task into two stages, generating deterministic and stochastic components and merging their useful information together. This takes advantage of different network architectures and optimizes the network output.

3.3 Image Fusion Network

As shown in Figure 2, our image fusion network can be decomposed into three parts: a kernel prediction module, a structure mapping module, and a structure-guided fusion module.

Kernel Prediction Module. Because the proportion of the stochastic component introduced into the final high-resolution image is not invariant when different stochastic weights are used, it is necessary to use a kernel prediction module to modulate the kernel weight and bias of our deterministic-stochastic fusion network. Inspired by MetaSR [15], our kernel prediction module includes a hidden layer of 16 neurons, and the stochastic weight α is taken as the input. Kernel weight and bias of one of the convolutional layers in our fusion network are prescribed by the kernel prediction module and are controlled by stochastic weight. After the training stage, the kernel prediction module can generate kernel weight and bias for the associated meta convolutional layer according to any given stochastic weight, and these predicted kernel parameters adjust our fusion network automatically without the need to re-train the network.

Structure Mapping Module. To make the fusion network aware of the spatial distribution of structural and non-structural pixels, we introduce a structure mapping module to infer a structure map. Because structural information belongs to low-level features, we use five cascaded convolutional layers as a structure mapping branch, which is similar to VGG16 before the second max-pooling layer. To expand the receptive field without the loss of resolution, we also remove the first max-pooling layer and replace the following conventional convolutional layer with a dilated convolutional layer. Although we do not assign a separate loss function to supervise the structure mapping module, experimental results have indicated that this module succeeds in capturing the structures in the input images.

Structure-guided Fusion Module. As shown in Figure 2, the structure-guided fusion module takes an image with a deterministic component, another image with a stochastic component, and a structure map generated by the structure mapping module as the input. To facilitate the convergence of the network, we first obtain a simple prediction of the output image by linearly interpolating the deterministic and stochastic components using α and then use the fusion module to infer the residual between the final output image and this simple prediction. Although we use only three convolutional layers in the fusion module, it is surprisingly effective to handle the input images. Kernel parameters for the middle convolutional layer in the fusion module are prescribed by the kernel prediction module. The structure of our proposed fusion network is a modified version of the basic residual unit from Reference [42].

3.4 Image Structure Labelling

It has been observed that the weights for deterministic and stochastic components are not evenly distributed across all pixels. Salient edges leave strong traces in low-resolution images, and the network for deterministic component generation is able to reconstruct them accurately. We denote these structural areas as M and use them to generate losses during training, so that the network will learn their distribution implicitly and do inference after the training. To avoid missing many possible structures, we use the commonly used Canny edge detector rather than other high-end, structure-specific boundary detectors such as COB [22], HED [44], or ELSD [27].

According to the analysis, if an edge shows up in $I^{sr,d}$ and I^{hr} simultaneously, then it is a deterministic structure. If an edge is detected at the same location in both $I^{sr,d}$ and I^{hr}, then any pixel on this edge segment is considered as a structural pixel. Because of possible localization error, a margin δ_i is calculated as the distance from the location of the current pixel to the closest edge pixel, and any pixels with distance δ_i smaller than a given threshold δ_a are considered as structural pixels. Let t_i denote one of the detected edge pixels $\mathcal{T}$ in the ground-truth image at location i and s_j denote one of the detected edge pixels $\mathcal{S}$ in the super-resolution image at location j; then the structure guide M_i at pixel location i is calculated as follows:

$$\begin{aligned} \delta_i &= \min_{s_j \in \mathcal{S}} ||t_i - s_j||^2, \\ M_i &= \mathcal{X}(\delta_i - \delta_a), \end{aligned} \tag{1}$$

where $\mathcal{X}(x) = 0$ when $x > 0$ and $\mathcal{X}(x) = 1$ otherwise. All pixels within the margin δ_a are considered as structural pixels. Pixels within the structural area are encouraged to align with the deterministic component. In our experiments, δ_a is set to 3. This whole process is shown in Figure 3.

3.5 Structure-modulated Loss

The problem with image interpolation is that although it guarantees a smooth transition from distortion-oriented results toward perception-oriented results when the stochastic weight is continuously increasing, it treats all pixels in the image alike. As a matter of fact, structural

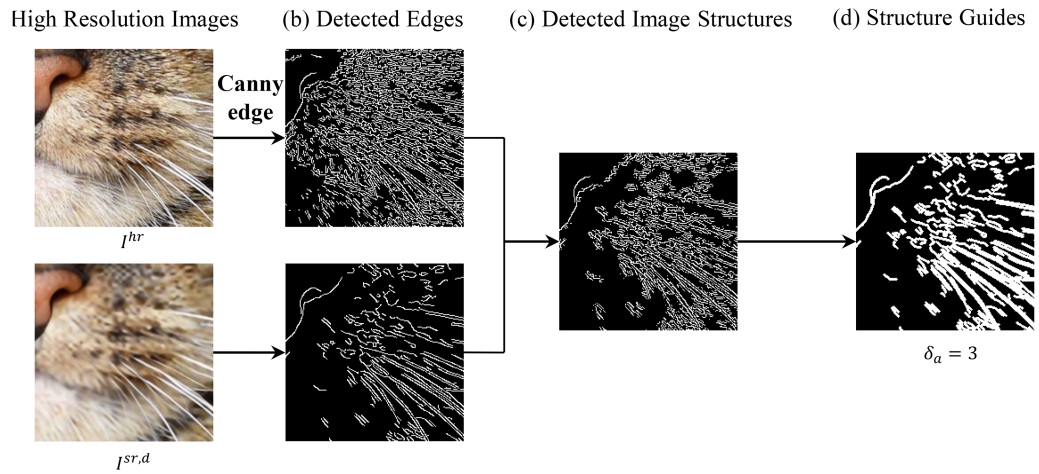

Fig. 3. Pipeline for image structure extraction. (a) High-resolution image I^{hr} and deterministic component $I^{sr,d}$. (b) Their Canny edges are extracted. (c) The intersection of their edges is extracted as the structure. (d) All pixels within the distances δ_a from extracted structures are structural pixels.

areas reconstructed by the respective networks for deterministic components and stochastic components are quite different. Linearly interpolating them does not take this problem into consideration and destroys the structures in both images.

Existing work on model interpolation [41] has taken the non-linearity of models into consideration, but it is still not guaranteed that the interpolated model performs optimally at every working point: Weights from some layers have more than 80% absolute differences even if the perception-oriented model is fine-tuned from the distortion-oriented model and might tremendously amplify the noise because of the high non-linearity of the model. Model interpolation [41] also requires the same backbone network for both interpolated models, which limits the flexibility of network design, as some network architectures could be more capable in generating deterministic/stochastic components than the others.

Stochastic components behave like uncorrelated signals compared with deterministic components; thus, they can be considered as addictive noise that we cannot predict. But removing them improves the distortion metrics of predicted super-resolution images. Because stochastic components are lost during down-sampling, it is impossible to reconstruct stochastic components accurately. However, stochastic components are crucial to human cognition, especially in areas without salient edges and other significant high-frequency information. Therefore, we rely on spatially varying interpolation between deterministic and stochastic components to introduce the stochastic component into the final high-resolution image. In particular, we only keep the deterministic component in structural areas and introduce the stochastic component into texture areas by interpolating the deterministic and stochastic components. This gives rise to the structure-modulated loss for training our fusion network,

$$\mathcal{L} = \sum_{m,n} \frac{M_{m,n}}{3n_0} ||I^{sr} - I^{sr,d}||_1 + \sum_{m,n} \frac{1 - M_{m,n}}{3n_1} ||I^{sr} - [I^{sr,d} + \alpha I^{sr,s}]||_1, \tag{2}$$

where M is the structure guide as described in Section 3.4, I^{sr} is the synthesized high-resolution image produced by our fusion network, $I^{sr,d}$ is the input image with the deterministic component, $I^{sr,s}$ is the input image with the stochastic component, n_0 denotes the number of structural

pixels, n_1 the number of non-structural pixels in the training image patch, and α is the stochastic weight, which is randomly sampled during training to obtain non-uniformly interpolated images with varying degrees of stochasticity as the supervision of our fusion network. This structure-modulated loss is the only loss function in our fusion network. Despite its simplicity, it can effectively train our fusion network to produce super-resolution images with varying degrees of stochasticity.

4 EXPERIMENTS

4.1 Training Details and Datasets

In the first stage, similarly to many perception-oriented models, we use a scaling parameter of $\times 4$ during down-sampling for the training of the backbone. The deterministic component is obtained with a pretrained ESRGAN backbone [42] with L1 loss. Two stochastic components are obtained with a pretrained ESRGAN backbone and a pretrained BOE backbone [26] with perceptual-oriented training settings described in their papers, respectively.

To introduce rich and diverse structures and textures in our training process, we follow the same setting in Reference [42], and our training data include the DIV2K training set [1], the OutdoorScene training set [40], and the Flickr2K dataset [33]. We evaluate our proposed methods on datasets with various attributes: Set5 [4], Set14 [48], B100 [23], Urban100 [16], Manga 109 [24], the DIV2K validation set [1], and the PIRM validation set [5].

In the second stage, the fusion network is trained with a mini-batch size of 16. The images are cropped into patches of size 128×128 during training. The image patches are also randomly flipped and rotated for simple data augmentation. We adopt the ADAM algorithm to do the training. The initial learning rate is set to 1×10^{-4} and exponential decay rates βs are $[0.9, 0.99]$, respectively. The training continued for $40k$ iterations. The stochastic weight α is generated randomly from 0 to 1 for each training sample. The structure margin δ_a is set to 3 during training. We conduct the training on one RTX2080Ti, and the training takes two days. The inference time for a 2K image pair is roughly 0.2 s.

4.2 Quantitative Results

We verify the performance of our proposed algorithm on commonly used super-resolution datasets under both blind image quality metric NIQE [25] and fully referenced metric RMSE. NIQE is a no-reference image quality assessment metric to estimate the perceptual quality of an image and is fast enough to evaluate a large number of images generated with different stochasticity levels in our experiments. The NIQE is trained on 125 pristine images with patch size set to 96×96 and sharpness threshold 0.75. To make a comparison with other perception–distortion controllable methods such as model interpolation [41], image interpolation [42], noise-tuned method BOE [26], and flow-based method [21], two meta-fusion networks with stochastic backbones of ESRGAN and BOE (track 3) are used respectively to make a comparison.

We evaluate these algorithms on the perception–distortion plane. Because it is impossible to evaluate every working point on the curves, we evaluate the model performance with several working points by selecting different stochastic parameter α ranging from 0 to 1 with a step of 0.1 and interpolate the rest of the curves. Methods like SRFlow [21] and BOE [26] have different starting points, because these two methods have their own backbone networks that are highly integrated into their method and cannot be altered. We find that given the suitable deterministic and stochastic backbones, our fusion network can produce the best results under both metrics of NIQE and RMSE on every dataset, as shown in Figure 4. It is not surprising to see that our proposed algorithm has achieved the most significant performance improvement over datasets

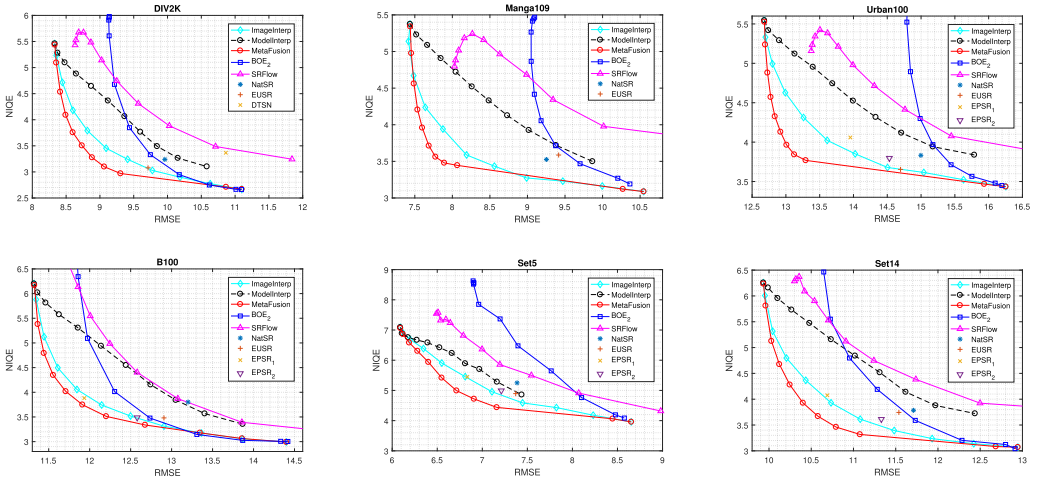

Fig. 4. Quantitative comparison of state-of-the-art methods on commonly seen SISR datasets, including Set5, Set14, B100, Urban100, Manga 109, and DIV2K-val.

with rich structures such as Manga109 and Urban100. For instance, when the RMSE is controlled at around 7.7, the NIQE of our proposed method is around 0.5 better than the second best algorithm on Manga109, and the NIQE of our proposed method is around 0.6 better than the second-best algorithm on Urban100 when the RMSE is controlled at around 13.2. Even in datasets that consist of mostly low-resolution texture-rich objects such as B100, our model is still able to have a margin around 0.2 when RMSE is controlled around 12. Both improvements in NIQE and RMSE illustrate that our fusion network has the capability to extract structure from all kinds of images, preserve their deterministic components, and hallucinate the texture areas efficiently.

Our proposed fusion network is also highly efficient regarding computational complexity. If the computation cost of a forward pass of a backbone network is C, then the computational complexity of our proposed model is O(2C) for the first inference, because most computational cost is due to the deterministic and stochastic backbones. When the input image remains the same while the level of stochasticity is adjusted, the inference time becomes almost negligible due to the tiny size of our fusion network. As a comparison, the complexity of model interpolation [42] is O(2C), because it requires the models to be interpolated first before inference. The complexity of BOE [26] and SRFlow [21] is O(C), because they run the whole network during every inference and for every α value.

4.3 Visual Results

Our proposed algorithm has a better performance when dealing with structural areas and effectively preserves deterministic components when given a target distortion level. We compare our algorithm with other controllable interpolation methods, namely image interpolation and model interpolation described in Reference [41], and noise-adding strategy described in Reference [26]. We use ESRGAN backbones for fair comparison. Each row in Figure 5–Figure 12 from left to right is the ground-truth image and the magnified patches from BOE, image interpolation, model interpolation, and ours, respectively. We present all images with comparable RMSE in each row, because it will be meaningless to compare the perceptual performance without constraining the distortion level. The RMSE for the corresponding super-resolved image is at the bottom of the patch.

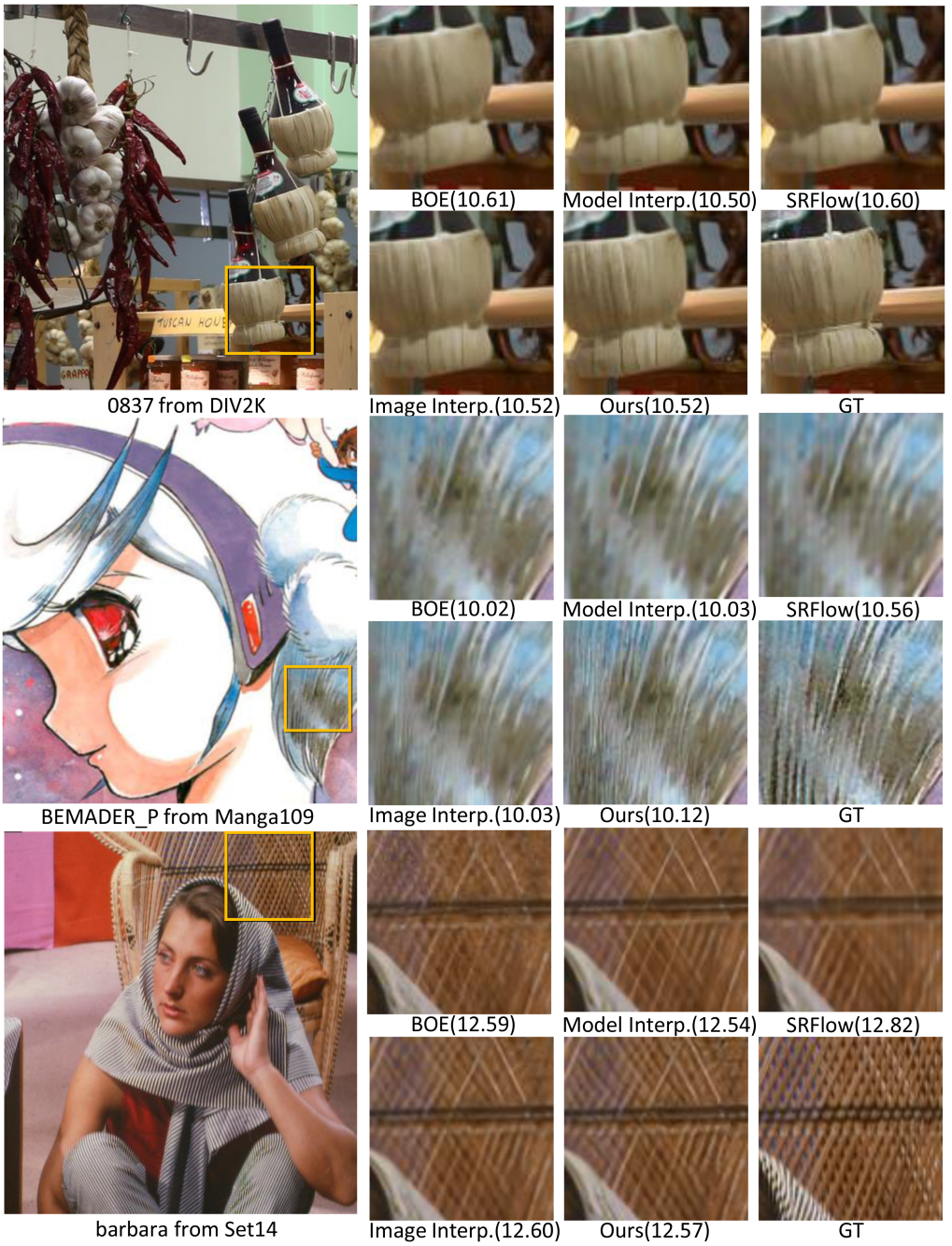

Fig. 5. Visual comparison of super-resolution images generated by different state-of-the-art fusion strategies such as BOE [26], model interpolation [42], and SRFlow [21] under similar RMSE (marked in (·)).

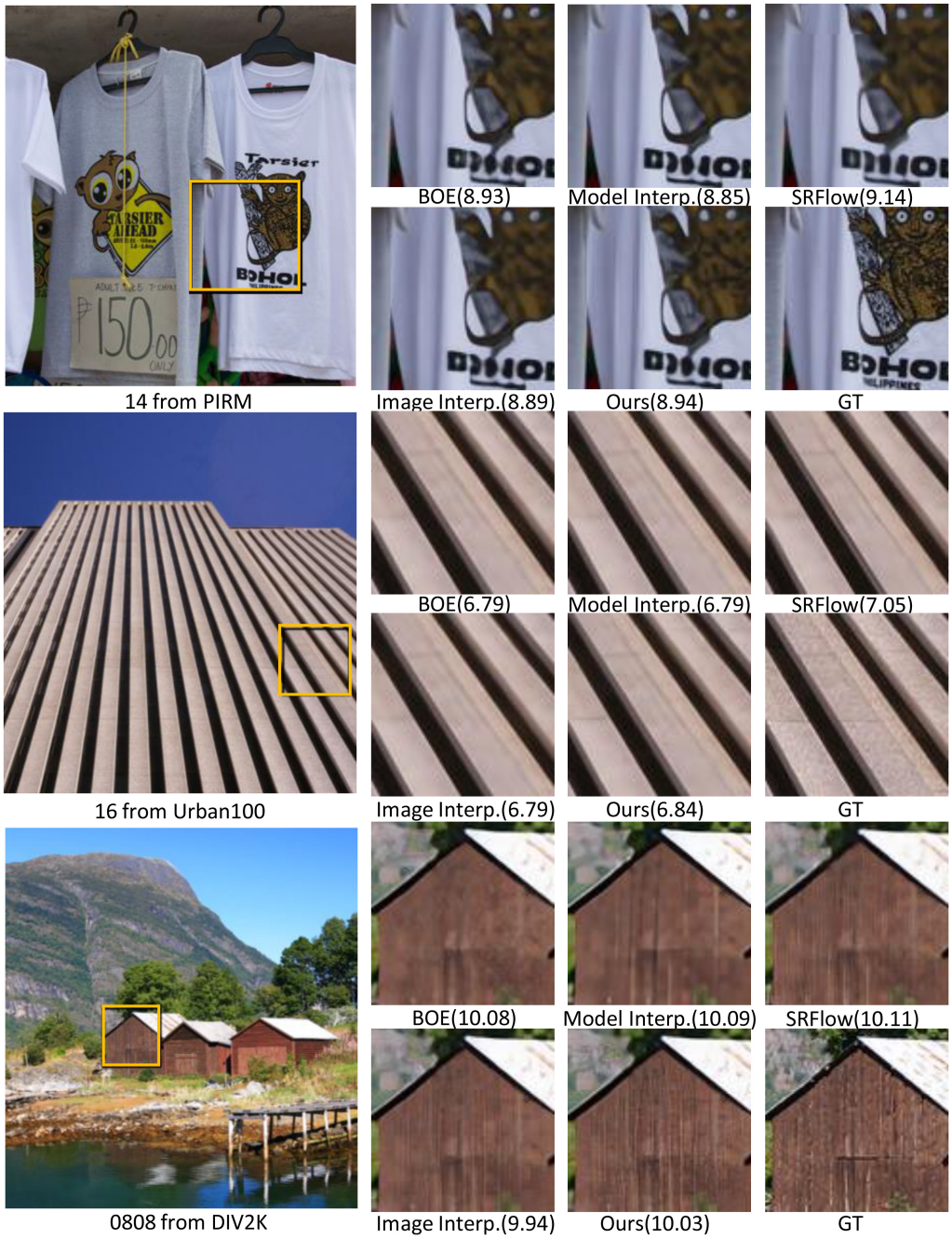

Fig. 6. Visual comparison of super-resolution images generated by different state-of-the-art fusion strategies such as BOE [26], model interpolation [42], and SRFlow [21] under similar RMSE (marked in (·)).

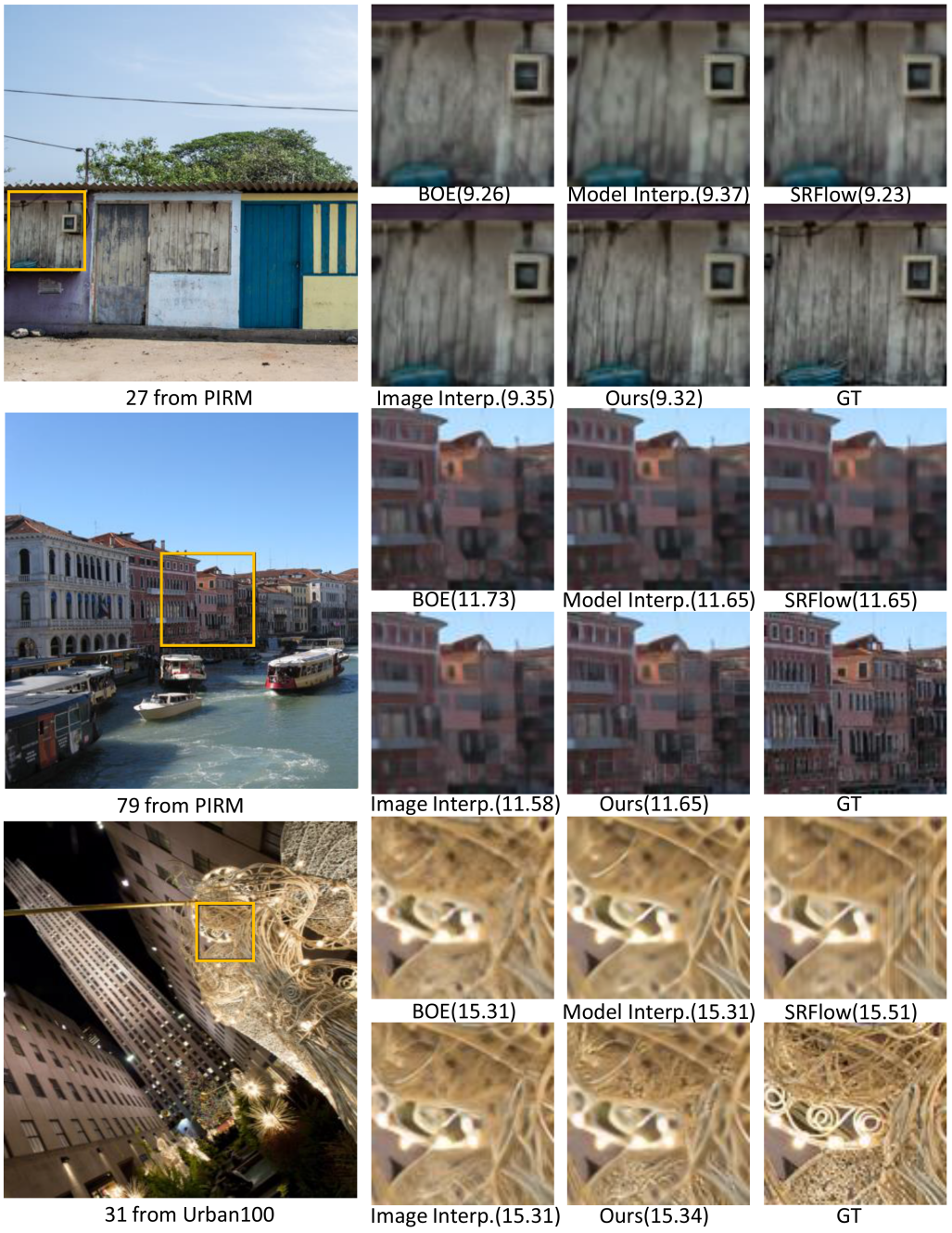

Fig. 7. Visual comparison of super-resolution images generated by different state-of-the-art fusion strategies such as BOE [26], model interpolation [42], and SRFlow [21] under similar RMSE (marked in (·)).

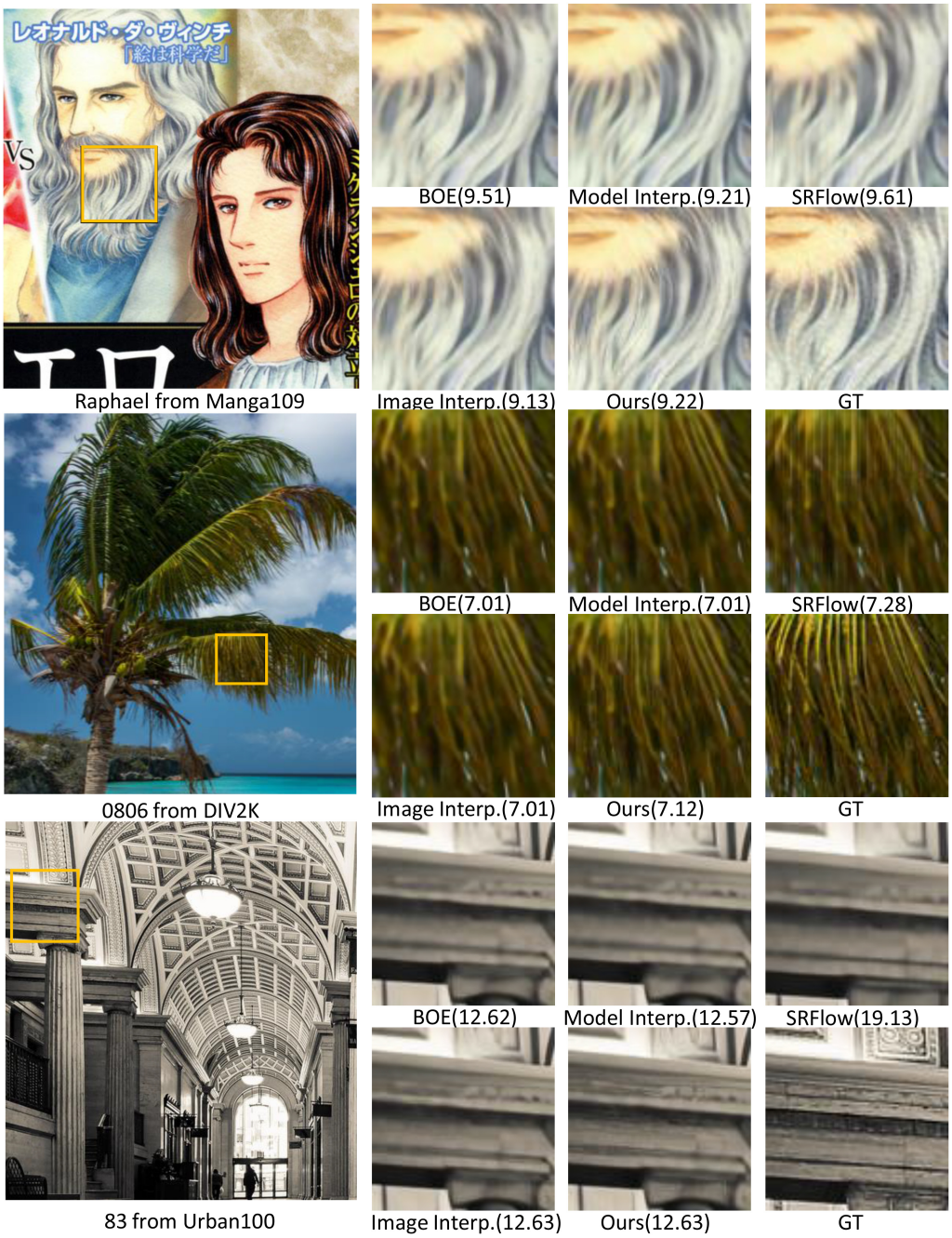

Fig. 8. Visual comparison of super-resolution images generated by different state-of-the-art fusion strategies such as BOE [26], model interpolation [42], and SRFlow [21] under similar RMSE (marked in (·)).

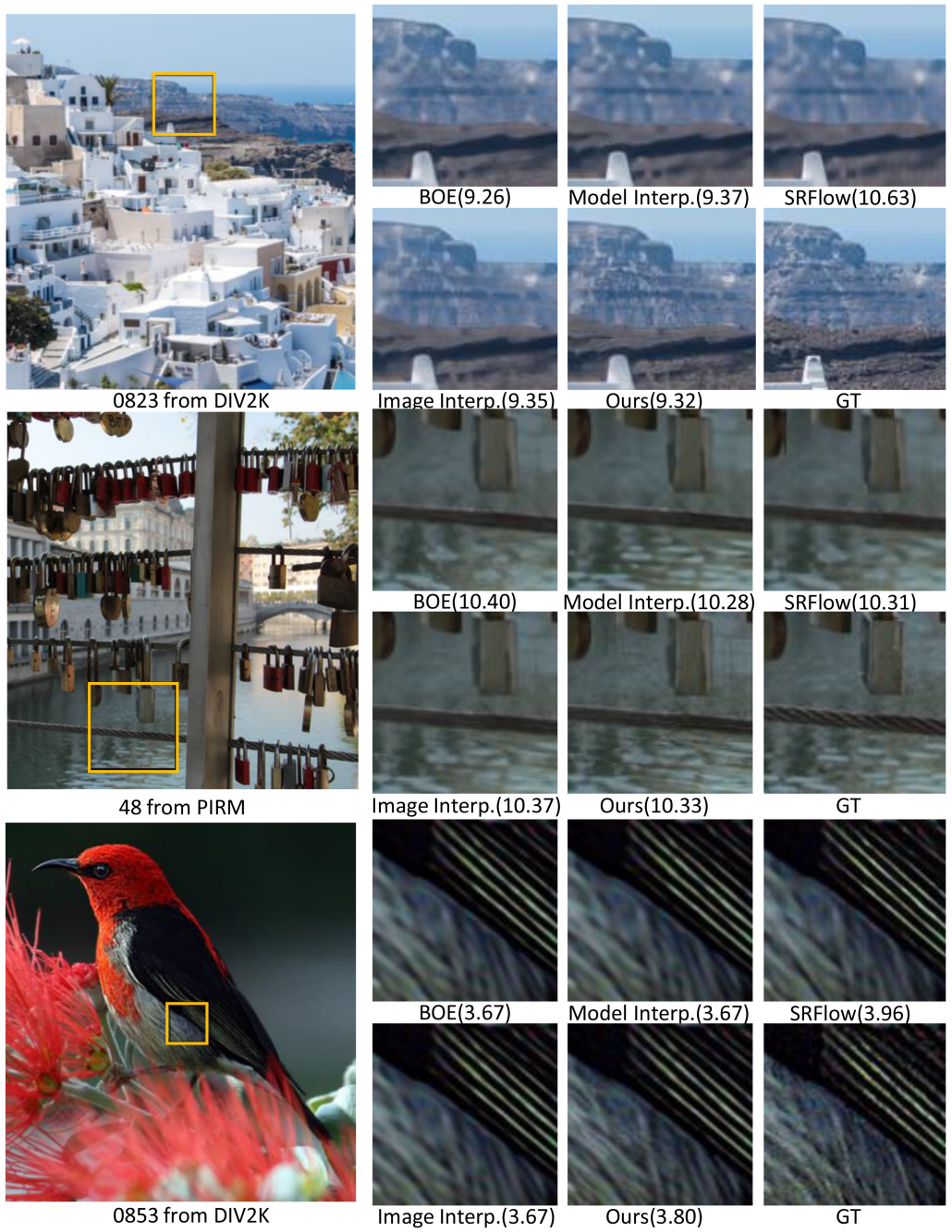

Fig. 9. Visual comparison of super-resolution images generated by different state-of-the-art fusion strategies such as BOE [26], model interpolation [42], and SRFlow [21] under similar RMSE (marked in (·)).

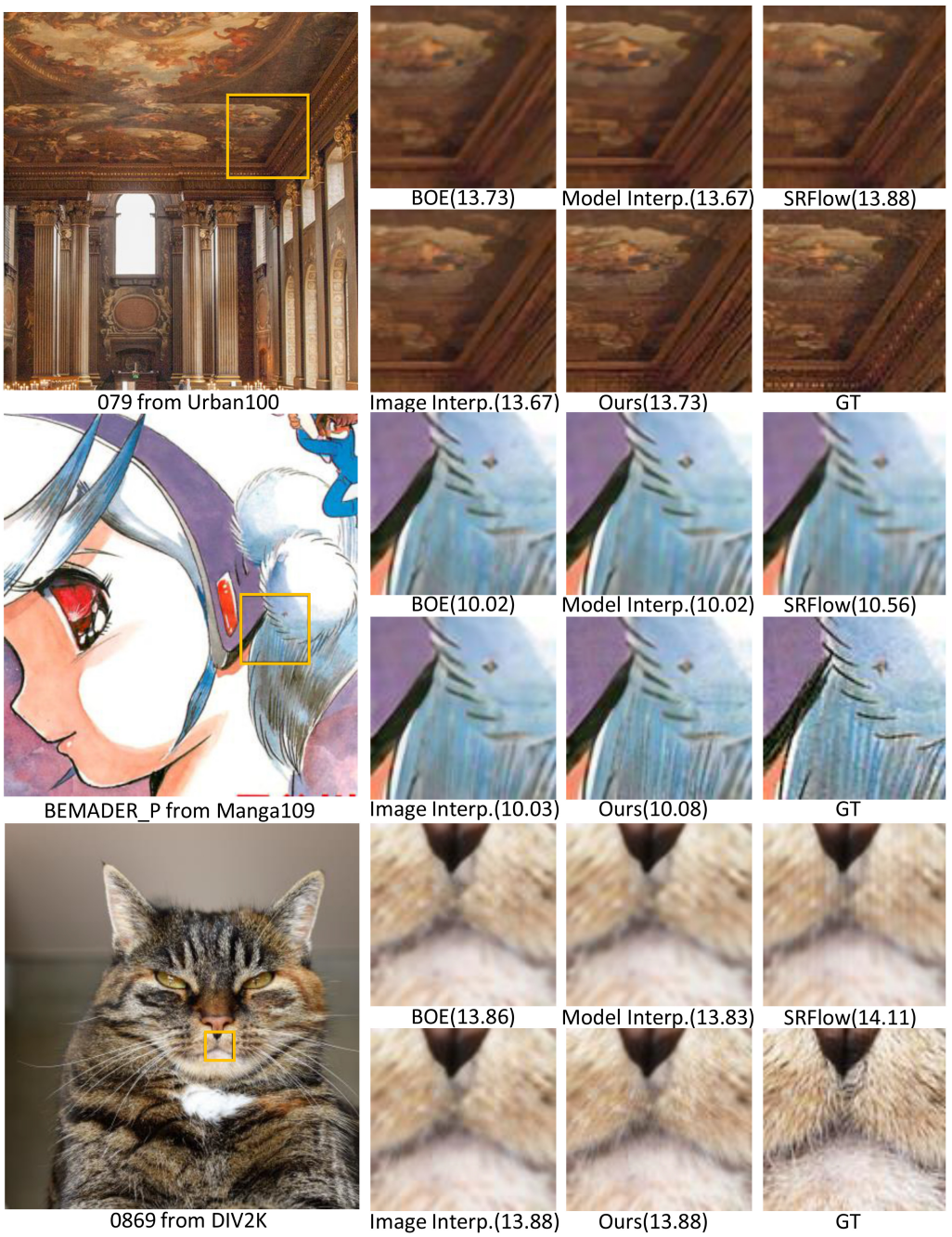

Fig. 10. Visual comparison of super-resolution images generated by different state-of-the-art fusion strategies such as BOE [26], model interpolation [42], and SRFlow [21] under similar RMSE (marked in (·)).

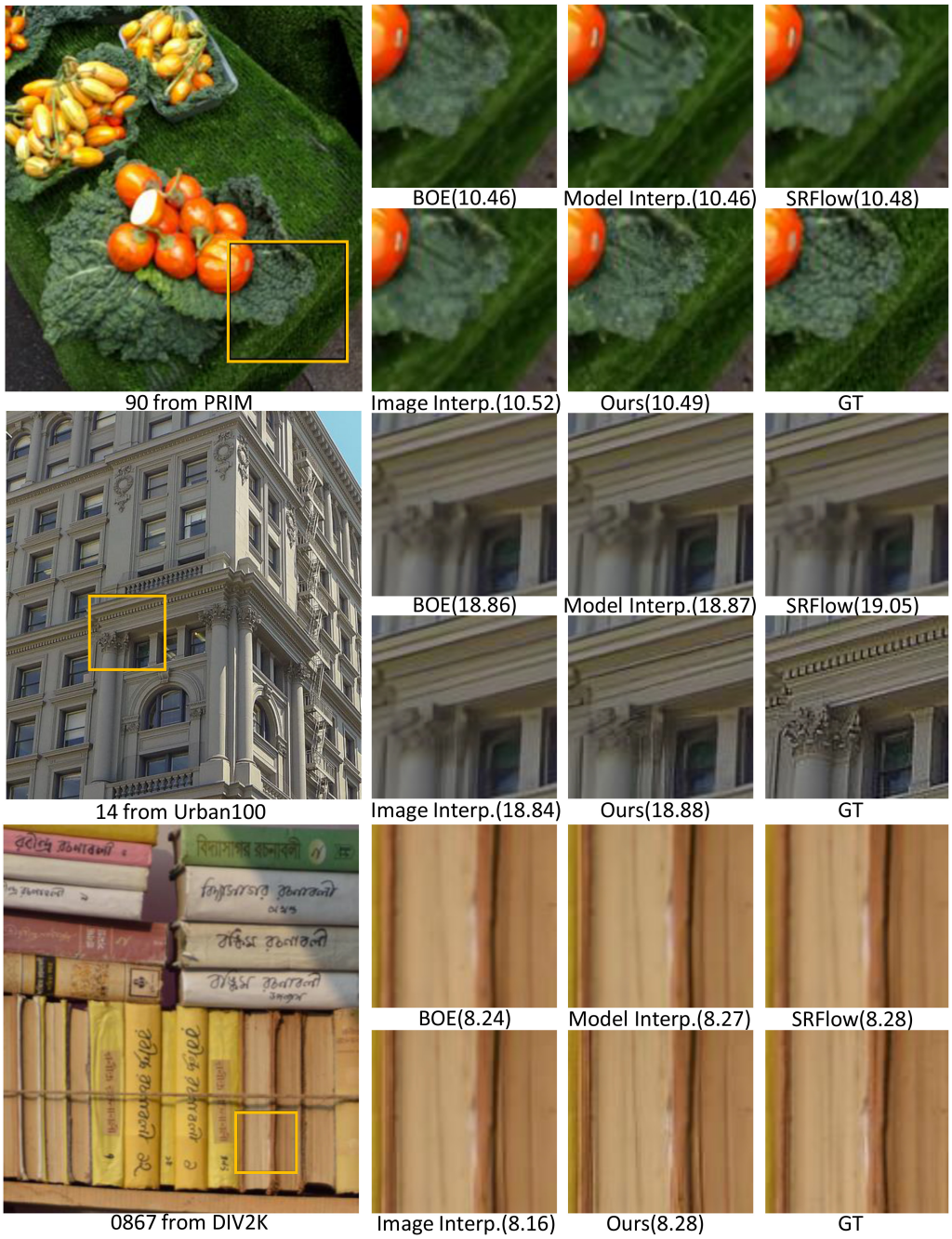

Fig. 11. Visual comparison of super-resolution images generated by different state-of-the-art fusion strategies such as BOE [26], model interpolation [42], and SRFlow [21] under similar RMSE (marked in (·)).

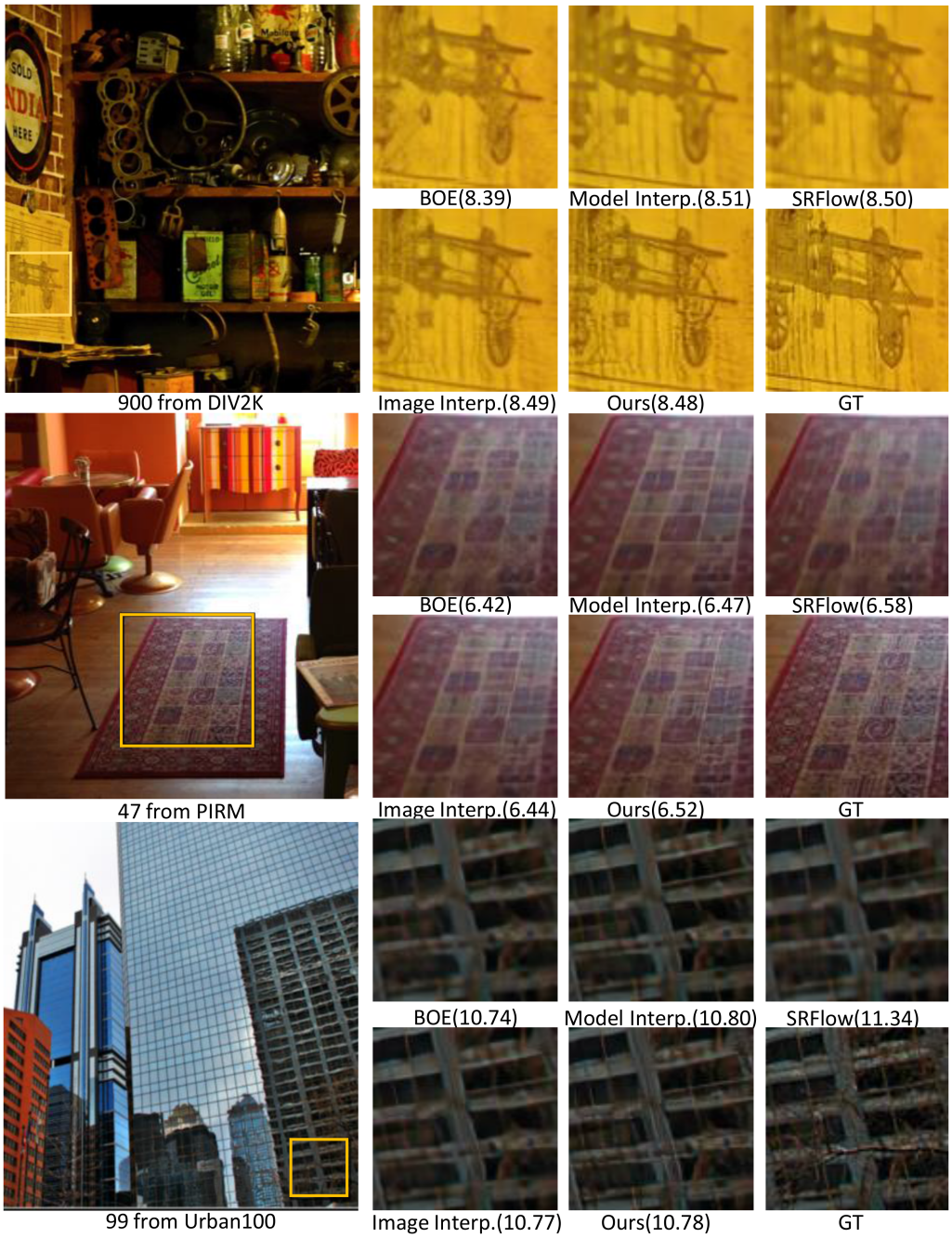

Fig. 12. Visual comparison of super-resolution images generated by different state-of-the-art fusion strategies such as BOE [26], model interpolation [42], and SRFlow [21] under similar RMSE (marked in (·)).

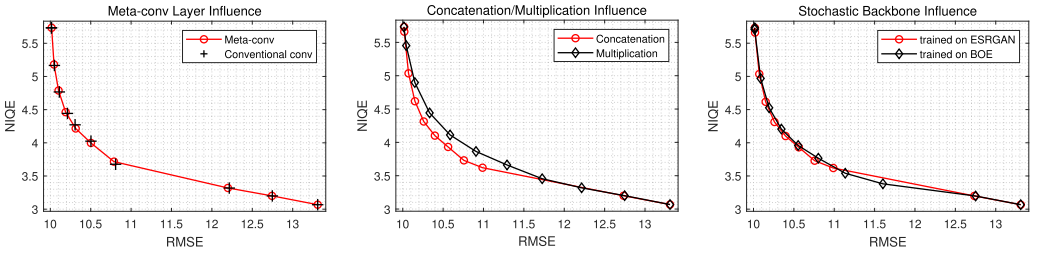

Fig. 13. Left: Kernel prediction module predicts convolutional kernels accurately. Middle: Concatenation performs better than multiplication in structure map. Right: Models are trained with different stochastic backbones but tested with the same stochastic backbone, which are almost identical.

Fig. 14. (a) High-resolution image I^{hr}. (b) Structure map (normalized for better visualization). (c) Structure guides.

It can be observed from Figure 5 that images generated by our method have better visual quality than other controllable interpolation methods. Given a fixed target distortion level, our method can produce more plausible stochastic details while preserving the accuracy of structures. This observation is also confirmed by the quantitative results from Figure 4. For instance, our methods can produce more accurate structures, such as building contours in images 14, 16, and 83 from Urban100. Moreover, our proposed method is also able to hallucinate more texture details on the back surface of barbara's chair, house wall in image 0808, and other high-frequency textures in 0823 and 0869 from DIV2K. Therefore, our method can produce more plausible stochastic details while preserving structure accuracy when a target distortion level is defined by the user. The advantage of our method exists at nearly all distortion levels and can be observed from images with different natures.

4.4 Ablation Studies

To investigate how the proposed modules influence the final performance of the algorithm, we conduct an ablation study by replacing the kernel prediction module. We also visualize the output of the structure map to evaluate its correlation with the given structure guide. To examine how provided stochastic components influence the final output image, we also test models trained with different stochastic backbones with the same stochastic inputs.

Kernel Prediction Module. By replacing the only meta-convolutional layer with a conventional convolutional layers and fixing the stochastic weight α, the model will lose the capability to adjust to a different α. But we will be able to examine how well the performance of our proposed algorithm is comparing to non-adjustable models. The result shown in Figure 13 (left) illustrates that our proposed method can predict the kernels trained separately under different stochastic weights α settings and provides almost identical results.

Structure Mapping module. We find that concatenation is better for convergence than multiplication to deal with the structure map, as shown in Figure 13 (middle). This might be because our mapping network are not supervised by the structure guide explicitly and the structure guide

itself is not a sparse signal. Although we do not apply any specific objective function to the structure mapping module, it can still successfully infer the image structure and pass it to the following structure-guided image fusion module. Several examples are shown in Figure 14. Comparing to image structure labels, it turns out most structures are extracted successfully even if we never use the structure label explicitly.

Although it is possible to use an edge detector to generate the structure map based on the $I^{sr,d}$, and use it for the deterministic edge-guided image interpolation, it is sub-optimal. This is because only edges both detected in the high-resolution and deterministic component are considered as correctly reconstructed structures, and edges detected only in the deterministic component are considered as falsely reconstructed structures. Since the high-resolution image is not available during inference, we will have to use a CNN to predict potential structure map. To verify the necessity of the structure map, we apply the same edge detection method to get the structure map based on the $I^{sr,d}$ and use it for the deterministic edge-guided image interpolation. The results are shown in Table 1. This result illustrates although deterministic edge-guided image interpolation performs better than direct image interpolation in most cases, it is stably worse than our proposed CNN-based method at any given level of RMSE. This is because the proposed CNN network can identify some falsely reconstructed edges in the deterministic component and alleviate the reconstruction error from them.

Moreover, the reconstruction of the deterministic component is not accurate. If we use edge detection methods to extract edge areas only in the deterministic component, then it will include falsely reconstructed edges/artefacts. Therefore, we use CNN to predict the structure map generated from both high-resolution images and the deterministic component is essential in our experiment.

Stochastic Backbone. Because the deterministic component is the least distorted information generated by the network, we expect structure areas that are mostly related to the deterministic component to avoid potential noise introduced by the stochastic backbone. To examine if the stochastic component undermines the generation of structure information, we train two fusion networks with the BOE stochastic backbone and the ERSGAN stochastic backbone, respectively, and test them on the same stochastic component generated by the ERSGAN backbone. If the stochastic backbone has little contribution in identifying structure areas, then changing them will not have a large difference on the final performance. The result shown in Figure 13 (right) demonstrates that they are almost identical, which verifies our expectation that structure information learned by the network is mostly extracted from deterministic component.

5 CONCLUSIONS

In this article, we have presented an image super-resolution framework that is capable of generating a high-resolution image by fusing a pair of deterministic and stochastic images using spatially varying weights. To make a single fusion model produce images with varying degrees of stochasticity, we further incorporate meta-learning into our fusion network in the form of a kernel prediction module. Experimental results indicate that our meta fusion network outperforms existing state-of-the-art SISR algorithms on widely used datasets, including PIRM-val, DIV2K-val, Set5, Set14, Urban100, Manga109, and B100.

APPENDICES

A EVALUATION METRICS

Here, we provide PSNR (RMSE) and NIQE results as in Reference [5], where RMSE is used to evaluate the distortion level and NIQE is used to evaluate the perceptual performance. PSNR is exactly

RMSE followed by a logarithmic operation. As discussed in Reference [5], SSIM is a metric for predicting the perceived quality; however, it has been increasingly criticized for its poor correlation with human perception of image quality. PSNR-HVS-M and PSNR-HVS [10, 28] take the Contrast Sensitivity Function and the between-coefficient contrast masking of DCT basis functions into evaluation, and mimic the human perception to some extent. These indices behave like the conventional distortion metric RMSE with a slight bias toward better perceptual performance, which makes these metrics inferior in judging the performance of the perceptual–distortion tradeof,f as they are considered a fixed combination of the two evaluation metrics. Therefore, we stick to our choices of evaluation metrics as our main metrics. Nevertheless, we also provide the evaluation results of these metrics. Results in Figure 15 illustrate that our proposed method is superior to all other tradeoff controllable methods such as ESRGAN [42], BOE [26], and a Flow-based method [21]. In addition, we achieve better or comparative performance with most non-controllable methods at certain tradeoff points over all testing datasets. But we stick to the same testing protocol as in Reference [5] according to the nature of SSIM and RMSE.

The results under PSNR-HVS-M and PSNR-HVS [5] are very similar, as shown in Figures 16 and 17.

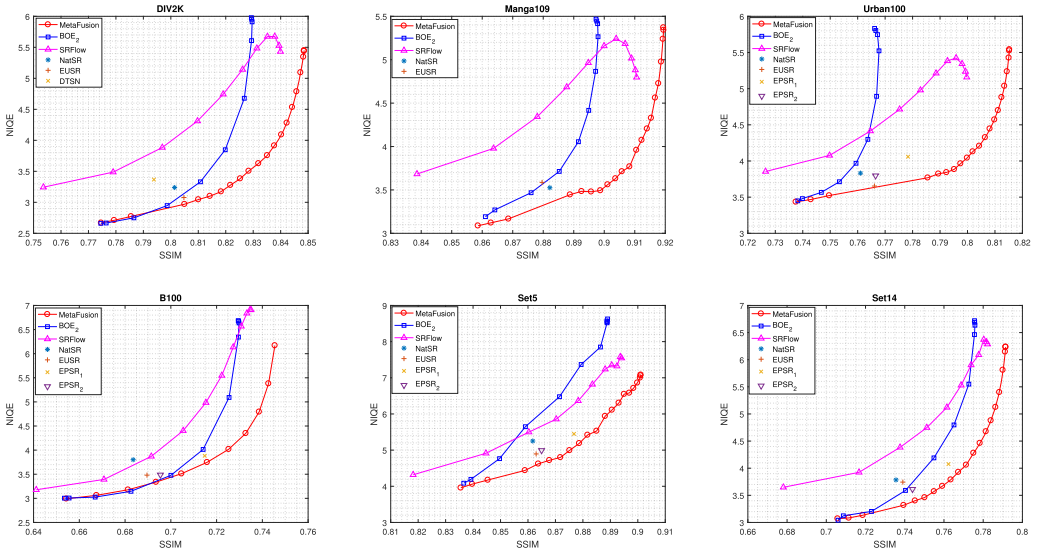

Fig. 15. Comparison of SSIM among state-of-the-art methods on commonly seen SISR datasets, including Set5, Set14, B100, Urban100, Manga 109, and DIV2K-val. Smaller NIQE and larger SSIM indicate better performance.

B STRUCTURAL TAGS

If we use traditional edge detection methods to extract edge areas in the deterministic component only, then it will include falsely reconstructed edges/artefacts. There are three types of edges: (a) edges detected in both the high-resolution image and deterministic component are correctly reconstructed structures; (b) edges detected in the high-resolution image only are considered as the true stochastic component, which is forever lost during down-sampling; and (c) edges detected in the deterministic component only are falsely reconstructed structures. The reconstruction of the deterministic component is not accurate. The structures defined in our experiments only refer to

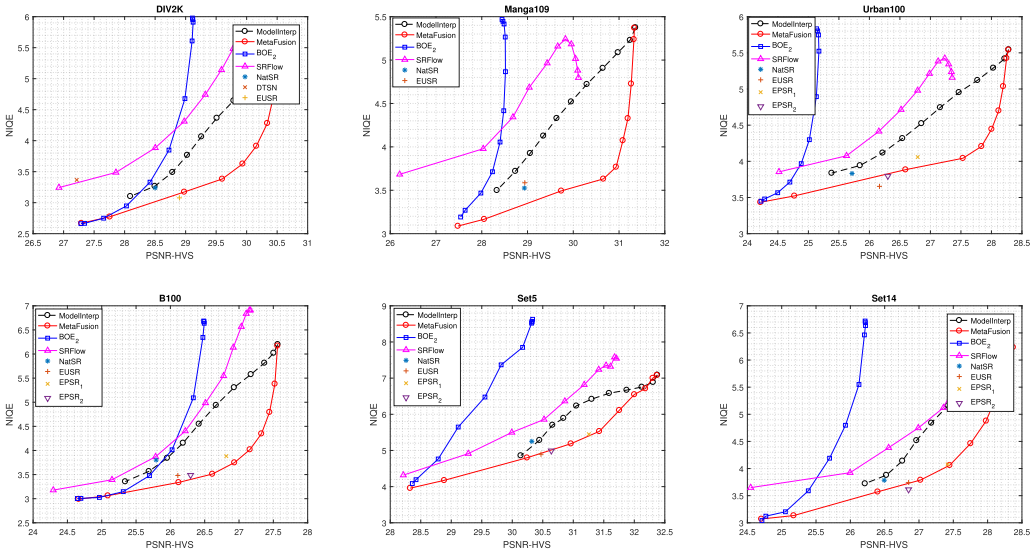

Fig. 16. Comparison of PSNR-HVS among state-of-the-art methods on commonly seen SISR datasets, including Set5, Set14, B100, Urban100, Manga 109, and DIV2K-val. Smaller NIQE and larger PSNR-HVS indicate better performance.

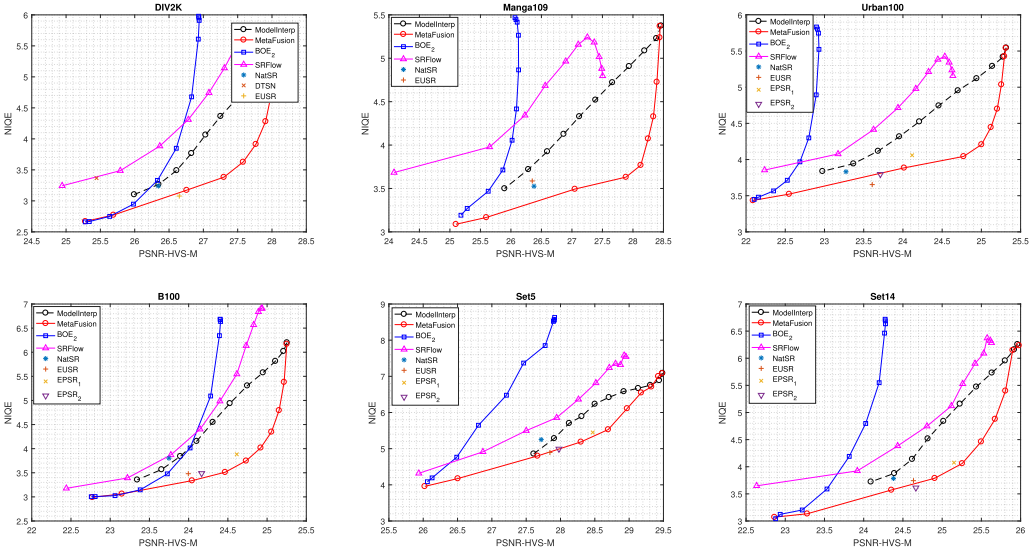

Fig. 17. Comparison of PSNR-HVS-M among state-of-the-art methods on commonly seen SISR datasets, including Set5, Set14, B100, Urban100, Manga 109, and DIV2K-val. Smaller NIQE and larger PSNR-HVS-M indicate better performance.

edges detected in both the high-resolution image and deterministic component. Since the high-resolution image is not available during inference, we have to use a CNN to classify potential structural areas. Therefore, using a CNN to predict the structure map generated from both the high-resolution image and the deterministic component is essential in our method.

Table 1. Evaluation of the Simple Image Interpolation, Deterministic Edge-guided Image Interpolation, and our Proposed Model on PIRM Validation Set

NIQE	10.0	10.3	10.6	10.9	11.2	11.5	11.8	12.1	12.4	12.6
image interp.	5.63	4.56	4.12	3.87	3.71	3.58	3.45	3.34	3.25	3.17
w/o high-res	5.60	4.34	4.01	3.79	3.64	3.52	3.44	3.37	3.29	3.22
proposed	**5.51**	**4.25**	**3.91**	**3.68**	**3.55**	**3.46**	**3.36**	**3.27**	**3.21**	**3.15**

Our proposed method always has a better perceptual index in comparison to deterministic edge-guided interpolation at any level of RMSE.

To illustrate this, our experiment uses a traditional edge detection method to generate the structural map from the deterministic HR output only and then uses this structural map to interpolate the deterministic and stochastic components by only keeping the deterministic component in the structural area while interpolating the two components in the non-structural area. The result is illustrated in Table 1. We test different strategies on PIRM validation set and follow the same evaluation metric as in the original paper. We compare their perceptual metric NIQE by controlling RMSE. Note that, generally, images with lower NIQE have better visual quality. The result indicates that images interpolated using deterministic edge maps are stably worse than our proposed CNN-based method at any given level of RMSE. This is because the proposed CNN network can identify falsely reconstructed edges in the deterministic component and decrease the reconstruction error caused by them.

REFERENCES

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

[2] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. 2016. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*. 3981–3989.

[3] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. 2018. Finding tiny faces in the wild with generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 21–30.

[4] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings British Machine Vision Conference*. 135.1–135.10.

[5] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. 2018. The 2018 PIRM challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 0–0.

[6] Yochai Blau and Tomer Michaeli. 2018. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6228–6237.

[7] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. 2004. Super-resolution through neighbor embedding. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Vol. 1. IEEE, I–I.

[8] Jun-Ho Choi, Jun-Hyuk Kim, Manri Cheon, and Jong-Seok Lee. 2020. Deep learning-based image super-resolution considering quantitative and perceptual quality. *Neurocomputing* 398 (2020), 347–359.

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2015. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 2 (2015), 295–307.

[10] Karen Egiazarian, Jaakko Astola, Nikolay Ponomarenko, Vladimir Lukin, Federica Battisti, and Marco Carli. 2006. New full-reference quality metrics based on HVS. In *Proceedings of the 2nd International Workshop on Video Processing and Quality Metrics*, Vol. 4.

[11] William T. Freeman, Thouis R. Jones, and Egon C. Pasztor. 2002. Example-based super-resolution. *IEEE Comput. Graph. Appl.* 22, 2 (2002), 56–65.

[12] Weifeng Ge, Bingchen Gong, and Yizhou Yu. 2018. Image super-resolution via deterministic-stochastic synthesis and local statistical rectification. *ACM Trans. Graph.* 37, 6 (2018), 1–14.

[13] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. 2019. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1604–1613.

[14] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. 2019. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3897–3906.

[15] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. 2019. Meta-SR: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1575–1584.

[16] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5197–5206.

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

[18] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. 2019. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10522–10531.

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

[20] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. 2020. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2359–2368.

[21] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. 2020. SRFlow: Learning the super-resolution space with normalizing flow. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*.

[22] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. 2016. Convolutional oriented boundaries. In *Proceedings of the European Conference on Computer Vision*. Springer, 580–596.

[23] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV'01)*, Vol. 2. IEEE, 416–423.

[24] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2017. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools Appl.* 76, 20 (2017), 21811–21838.

[25] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2012. Making a "completely blind" image quality analyzer. *IEEE Sign. Process. Lett.* 20, 3 (2012), 209–212.

[26] Pablo Navarrete Michelini, Dan Zhu, and Hanwen Liu. 2018. Multi–scale recursive and perception–distortion controllable image super–resolution. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 0–0.

[27] Viorica Pătrăucean, Pierre Gurdjos, and Rafael Grompone Von Gioi. 2012. A parameterless line segment and elliptical arc detector with enhanced ellipse fitting. In *Proceedings of the European Conference on Computer Vision*. Springer, 572–585.

[28] Nikolay Ponomarenko, Flavia Silvestri, Karen Egiazarian, Marco Carli, Jaakko Astola, and Vladimir Lukin. 2007. On between-coefficient contrast masking of DCT basis functions. In *Proceedings of the 3rd International Workshop on Video Processing and Quality Metrics*, Vol. 4.

[29] Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*.

[30] Jae Woong Soh, Gu Yong Park, Junho Jo, and Nam Ik Cho. 2019. Natural and realistic single image super-resolution with explicit natural manifold discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8122–8131.

[31] Jian Sun, Zongben Xu, and Heung-Yeung Shum. 2008. Image super-resolution using gradient profile prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[32] Yu-Wing Tai, Shuaicheng Liu, Michael S. Brown, and Stephen Lin. 2010. Super resolution using edge prior and single image detail synthesis. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2400–2407.

[33] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 114–125.

[34] Radu Timofte, Vincent De Smet, and Luc Van Gool. 2013. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*. 1920–1927.

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

[36] Subeesh Vasu, Nimisha Thekke Madam, and A. N. Rajagopalan. 2018. Analyzing perception-distortion tradeoff using enhanced perceptual super-resolution network. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 0–0.

[37] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. 2021. Exploring sparsity in image super-resolution for efficient inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4917–4926.

[38] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. 2021. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10581–10590.

[39] Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. 2021. Learning a single network for scale-arbitrary super-resolution. In *Proceedings of the IEEE Conference on International Conference on Computer Vision*. 10581–10590.

[40] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 606–615.

[41] Xintao Wang, Ke Yu, Chao Dong, Xiaoou Tang, and Chen Change Loy. 2019. Deep network interpolation for continuous imagery effect transition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1692–1701.

[42] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*.

[43] Yu-Xiong Wang and Martial Hebert. 2016. Learning to learn: Model regression networks for easy small sample learning. In *Proceedings of the European Conference on Computer Vision*. Springer, 616–634.

[44] Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 1395–1403.

[45] Xuan Xu and Xin Li. 2019. Scan: Spatial color attention networks for real single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[46] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. 2008. Image super-resolution as sparse representation of raw image patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[47] Deniz Yıldırım and Oğuz Güngör. 2012. A novel image fusion method using IKONOS satellite images. *J. Geodesy Geoinf.* 1, 1 (2012), 75–83.

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

[49] Kai Zhang, Luc Van Gool, and Radu Timofte. 2020. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3217–3226.

[50] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 286–301.

[51] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. 2019. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7982–7991.

Received December 2020; revised July 2021; accepted July 2021