



# Deep learning-based image super-resolution considering quantitative and perceptual quality

Jun-Ho Choi, Jun-Hyuk Kim, Manri Cheon, Jong-Seok Lee\*

School of Integrated Technology, Yonsei University, 85 Songdogwahak-ro, Yeonsu-gu, Incheon, Korea

## ARTICLE INFO

### Article history:

Received 4 September 2018

Revised 11 June 2019

Accepted 17 June 2019

Available online 21 October 2019

### Keywords:

Perceptual super-resolution

Deep learning

Aesthetics

Image quality

## ABSTRACT

Recently, it has been shown that in super-resolution, there exists a tradeoff relationship between the quantitative and perceptual quality of super-resolved images, which correspond to the similarity to the ground-truth images and the naturalness, respectively. In this paper, we propose a novel super-resolution method that can improve the perceptual quality of the upscaled images while preserving the conventional quantitative performance. The proposed method employs a deep network for multi-pass upscaling in company with a discriminator network and two qualitative score predictor networks. Experimental results demonstrate that the proposed method achieves a good balance of the quantitative and perceptual quality, showing more satisfactory results than existing methods.

© 2019 Elsevier B.V. All rights reserved.

## 1. Introduction

Single-image super-resolution, which is a task to increase the spatial resolution of low-resolution images, has been widely studied in recent decades. One of the simple solutions for the task is to employ interpolation methods such as nearest-neighbor and bicubic upsampling. However, their outputs are largely blurry because fine details of the images cannot be recovered. Therefore, many researchers have investigated how to effectively restore high-frequency details. Nevertheless, it is still highly challenging due to the lack of information in the low-resolution images, i.e., an ill-posed problem [1].

Until the mid-2010s, feature extraction-based methods have been proposed, including sparse coding [2], neighbor embedding [3], and Bayes forest [4]. After that, the emergence of deep learning for visual representation [5], which is triggered by an image classification challenge (i.e., ImageNet) [6], has also flowed into the field of super-resolution [7]. For instance, the super-resolution convolutional neural network (SRCNN) model proposed by Dong et al. [8] introduced convolutional layers and showed better performance than the previous methods.

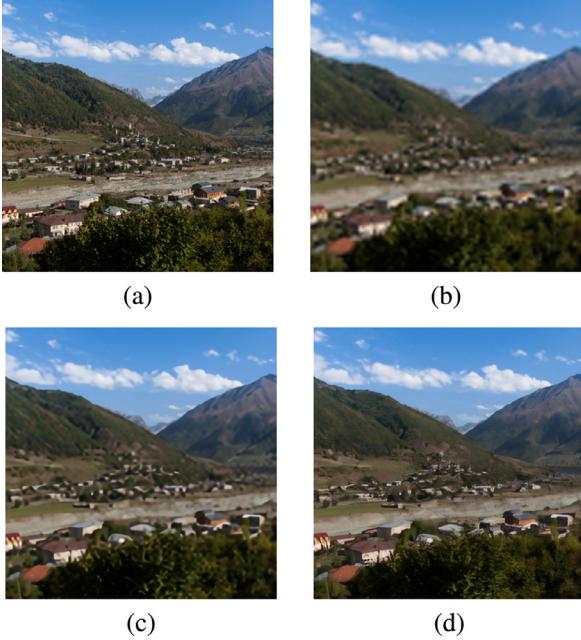
To build a deep learning-based super-resolution model, it is required to define loss functions that are the objectives of the model to be trained. Loss functions measuring pixel-by-pixel dif-

ferences of the ground-truth and upscaled images are frequently considered, including mean squared error and mean absolute error [7]. They mainly aim at guaranteeing quantitative conditions of the obtained images, which can be evaluated by quantitative quality measures such as peak signal-to-noise ratio (PSNR), root mean squared error (RMSE), and structural similarity (SSIM) [9]. Fig. 1(c) shows an example image generated by a deep learning-based super-resolution model, enhanced upscaling super-resolution (EUSR) [10], from the downsampled version of Fig. 1(a). Compared to the image upscaled by bicubic interpolation shown in Fig. 1(b), the image generated by the deep learning-based method follows the overall appearance of the original image with sharper boundaries of the objects and scenery.

Although existing methods based on minimizing pixel-by-pixel differences achieve great performance in a quantitative viewpoint, they do not ensure *naturalness* of the output images. For example, fine details of trees and houses are not sufficiently recovered in Fig. 1(c). To improve the naturalness of the images, two approaches have been proposed in the literature: using generative adversarial networks (GANs) [12] and employing intermediate features of the common image classification network models. For example, Ledig et al. [1] proposed a super-resolution model named SRGAN, which employs a discriminator network and trains the model to minimize differences of the intermediate features of VGG19 [13] when the ground-truth and upscaled images are inputted. It is known that these methods enhance perceptual performance significantly [14]. Here, the perceptual performance can be measured by the metrics for visual quality assessment such as blind/referenceless image spatial quality evaluator (BRISQUE) [15] and naturalness image quality evaluator (NIQE) [16].

\* Corresponding author.

E-mail addresses: [iidealbosome@yonsei.ac.kr](mailto:iidealbosome@yonsei.ac.kr) (J.-H. Choi), [junhyuk.kim@yonsei.ac.kr](mailto:junhyuk.kim@yonsei.ac.kr) (J.-H. Kim), [manri.cheon@yonsei.ac.kr](mailto:manri.cheon@yonsei.ac.kr) (M. Cheon), [jong-seok.lee@yonsei.ac.kr](mailto:jong-seok.lee@yonsei.ac.kr) (J.-S. Lee).



**Fig. 1.** Example results obtained for an image of the PIRM dataset [11]. (a) Ground-truth (b) upscaled by bicubic interpolation (c) upscaled without perceptual consideration (d) upscaled with perceptual consideration.

However, two issues still remain unresolved in these approaches. First, although these approaches improve naturalness of the images, perceptual quality is only indirectly considered and thus the improvement may be limited. The network models for extracting intermediate features are for image classification tasks, thus forcing the features to be similar does not guarantee perceptually improved results. In addition, it is possible that the discriminator network learns the criteria that can differentiate generated images from the real ones but are not related to the perceptual aspects. For instance, when the trained discriminator relies on just finding high-frequency components, the super-resolution model may add some unexpected textures in low-frequency regions such as ground and sky.

Second, these approaches tend to sacrifice a large amount of the quantitative quality. For example, the SRGAN-based models achieve better perceptual performance than the other models in terms of BRISQUE and NIQE, but they record worse quantitative quality, showing larger RMSE values [14]. Since the primary objective of the super-resolution task is to make the upsampled images identical to the ground-truth high-resolution images, it is necessary to properly regularize the upscaling modules to keep balance of the quantitative and qualitative quality.

In this paper, we propose a novel super-resolution method named “Four-pass perceptual super-resolution with enhanced upscaling (4PP-EUSR),” which is based on the recently proposed EUSR model [10]. Our model aims at resolving the aforementioned issues via two innovative ways. First, our model employs so-called “multi-pass upscaling” during the training phase, where multiple upsampled images produced by passing the given low-resolution image through the multiple upscaling paths in our model are used in order to consider various possible characteristics of upsampled images. Second, we employ qualitative score predictors, which directly evaluate the aesthetic and subjective quality scores of the upsampled images. This architecture ensures high perceptual quality with preserving the quantitative performance of the upsampled images, as exemplified in Fig. 1(d).

The rest of the paper is organized as follows. First, we provide a brief review of the related work in Section 2. Then, an overview

of the proposed method is given in Section 3, including the base deep learning model, multi-pass upscaling for training, structure of the discriminator, and structures of the qualitative score predictors. We explain training procedures of our model with the employed loss functions in Section 4. In-depth experimental analysis of our results is shown in Section 5. Finally, we conclude our work in Section 6.

## 2. Related work

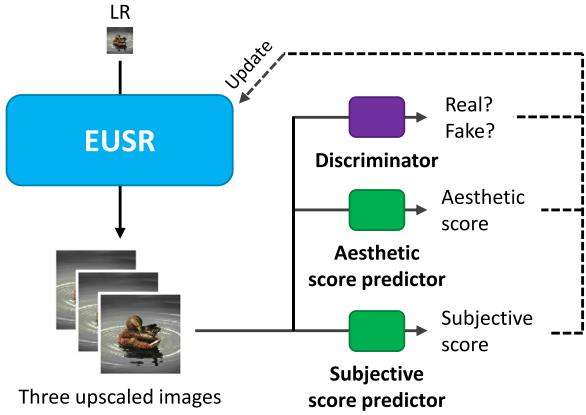
In this section, we review the related work of deep learning-based super-resolution in two branches: super-resolution models without and with consideration of naturalness.

### 2.1. Deep learning-based super-resolution

One of the earliest super-resolution models based on deep learning is SRCNN, which was proposed by Dong et al. [8]. The model takes an image upsampled by the bicubic interpolation and enhances it via two convolutional layers. Kim et al. proposed the very deep super-resolution (VDSR) model [17], which consists of 20 convolutional layers. In recent days, residual blocks having shortcut connections [18] are commonly used in the super-resolution models. For example, Ledig et al. [1] proposed a model named SRResNet, which contains 16 residual blocks with batch normalization [19] and parametric ReLU activation [20]. Lim et al. [21] developed two super-resolution models for the NTIRE 2017 single-image super-resolution challenge [22]: the enhanced deep super-resolution (EDSR) model for single-scale super-resolution and the multi-scale deep super-resolution (MDSR) model for multi-scale super-resolution. They found that removing batch normalization and blending outputs generated from geometrically transformed inputs help improving the overall quantitative quality. Recently, Kim and Lee [10] suggested a multi-scale super-resolution method named EUSR, which consists of so-called “enhanced upscaling modules” and performed well in the NTIRE 2018 single-image super-resolution challenge [23]. Zhang et al. [24] proposed a super-resolution model based on residual dense network (RDN), which extends the residual network to have densely-connected layers. Zhang et al. [25] proposed a residual channel attention networks (RCAN), which brings an attention mechanism into the super-resolution task and achieves better quantitative performance than EDSR.

### 2.2. Super-resolution considering naturalness

Along with ensuring high quantitative quality in terms of PSNR, RMSE, or SSIM, naturalness of the upsampled images, which can be measured by quality metrics such as BRISQUE and NIQE, has been also considered in some studies. There exist two common approaches: employing GANs [12] and employing image classifiers. In the former approach, the discriminator network tries to distinguish the ground-truth images from the upsampled images and the super-resolution model is trained to fool the discriminator so that it cannot distinguish the upsampled images properly. When an image classifier is used, the super-resolution model is trained to minimize the difference of the features obtained at the intermediate layers of the classifier for the ground-truth and upsampled images. For example, Johnson et al. [26] used the trained VGG16 network to extract the intermediate features and regarded the squared Euclidean distance between them as the loss function. Ledig et al. [1] employed an adversarial network and differences of the features obtained from the trained VGG19 network for calculating losses of their super-resolution model (i.e., SRResNet), which is named as SRGAN. Mechrez et al. [27] defined the so-called “contextual loss,”



**Fig. 2.** Overview of the proposed method. First, our super-resolution model (Section 3.1) generates three upscaled images via multi-pass upscaling (Section 3.2). The discriminator tries to differentiate the upscaled images from the ground-truth (Section 3.3). The two qualitative score predictors measure the aesthetic and subjective quality scores, respectively (Section 3.4). The outputs of the discriminator and the score predictors are used to update the super-resolution model.

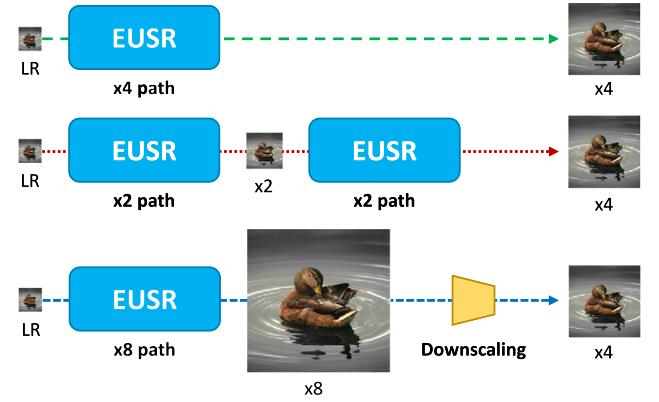
which compares the statistical distribution of the intermediate features obtained from the trained VGG19 model, to train their super-resolution model. These models focus on ensuring naturalness of the upscaled images but tend to sacrifice a large amount of the quantitative quality [14].

### 3. Overview of the proposed method

The architecture of the proposed method can be disassembled into four components (Fig. 2): a multi-scale upscaling model, employing the model in a multi-pass manner, a discriminator, and qualitative score predictors.

#### 3.1. Enhanced upscaling super-resolution

The basic structure of our model is from the EUSR model [10], which is shown in Fig. 3. It mainly consists of three parts: scale-aware feature extraction, shared feature extraction, and enhanced upscaling. First, the scale-aware feature extraction part extracts low-level features from the input image by using so-called “local residual blocks.” Then, a residual module in the shared feature extraction part, which consists of local residual blocks and a convolutional layer, extracts higher-level features regardless of the scale factor. Finally, the proceeded features are upscaled via “enhanced upscaling modules,” where each module increases the spatial resolution of the input by a factor of 2. Thus, the  $\times 2$ ,  $\times 4$ , and



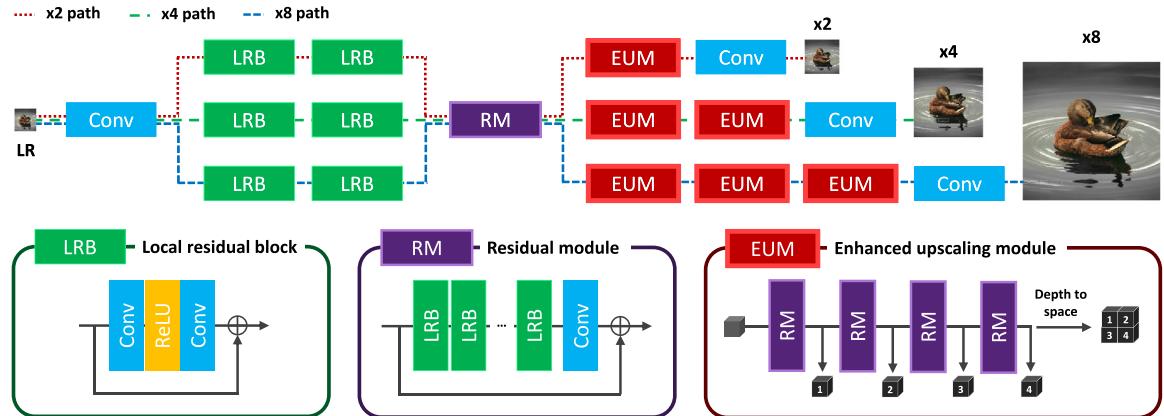
**Fig. 4.** Multi-pass upscaling process, which produces three upscaled images by a factor of 4 from a shared pre-trained EUSR model.

$\times 8$  upscaling paths have one, two, and three enhanced upscaling modules, respectively. The configurable parameters of the EUSR model are the number of output channels of the first convolutional layer, the number of local residual blocks in the shared feature extraction part, and the number of local residual blocks in the enhanced upscaling modules. We consider EUSR as our base upscaling model because it is one of the state-of-the-art approaches supporting multi-scale super-resolution, which enables generating multiple upscaled images from a single model.

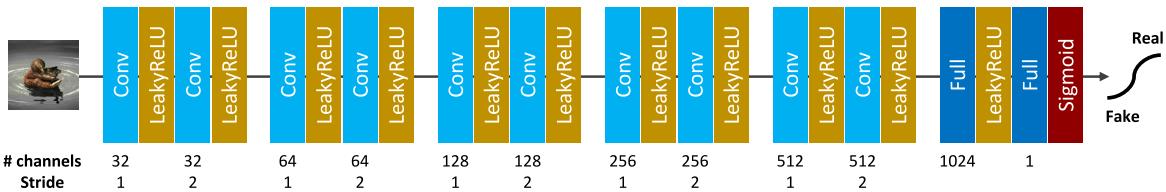
#### 3.2. Multi-pass upscaling

The original EUSR model supports multi-scale super-resolution by factors of 2, 4, and 8. During the training phase, our model utilizes all these upscaling paths to produce three output images, where we make the output images have the same upscaling factor of 4 for a given image as follows (Fig. 4). The first one is directly generated from the  $\times 4$  path. The second one is generated by passing the given image through the  $\times 2$  path two times. The third one is generated via bicubic downscaling of the image obtained from the  $\times 8$  path by a factor of 2. Thus, the model is employed four times for each input image.

The original purpose of multi-scale models such as MDSR [21] and EUSR [10] is to support variable scaling factors on a single model. On the other hand, our multi-pass upscaling extends it with a different objective, which is to improve the quality of the upscaled images for a fixed scaling factor. Since all three images obtained from different upscaling paths are used for training, the model has to learn reducing artifacts that may occur during di-



**Fig. 3.** Structure of the EUSR model [10].



**Fig. 5.** Structure of the discriminator network.

rect upscaling via the  $\times 4$  path, two-pass upscaling via the  $\times 2$  path, and upscaling via the  $\times 8$  path and downscaling. This prevents the model to overfit towards specific patterns, thus it enables the model to handle various upscaling scenarios.

### 3.3. Discriminator network

Our method employs a discriminator network during the training phase, which is designed to distinguish generated images from the ground-truth images. While the discriminator tries to do its best for identifying the upscaled images, the super-resolution model is trained to make the discriminator difficult to differentiate them from the ground-truth images. This helps our upscaling model generating more natural images [1,27]. Inspired by SRGAN [1], our discriminator network consists of several convolutional layers followed by LeakyReLU activations with  $\alpha = 0.2$  and two fully-connected layers, as shown in Fig. 5. The final sigmoid activation determines the probability that the input image is real or fake. Note that our discriminator network does not employ the batch normalization [19], because the batch size is too small to use that optimization. In addition, it contains two more convolutional layers than the original SRGAN model due to the different size of the input image patches.

### 3.4. Qualitative score predictors

One of our main ideas for perceptually improved super-resolution is to utilize deep learning models classifying perceptual quality of images, instead of general image classifiers. For this, we employ two deep networks that predict aesthetic and subjective quality scores of images, respectively. To build the networks, we utilize the neural image assessment (NIMA) approach [28], which predicts the quality score of a given image. This approach replaces the last layer of a well-known image classifier such as VGG [13] or Inception-v3 [29] with a fully-connected layer with the softmax activation, which produces probabilities of 10 score classes. In our implementation, MobileNetV2 [30] is used as the base image classifier, because it is much faster than the other image classifiers and supports various sizes of input images.

We build two score predictors: one for predicting *aesthetic* scores and the other for predicting *subjective* scores. For the aesthetic score predictor, we employ the AVA dataset [31], which contains aesthetic user ratings of the images shared in DPChallenge<sup>1</sup>. For the subjective score predictor, we use the TID2013 dataset [32], which consists of the subjective quality evaluation results for the test images degraded by various distortion types (e.g., compression, noise, and blurring). While the AVA dataset provides exact score distributions, the TID2013 dataset only provides the mean and standard deviation of the scores. Therefore, we approximate a Gaussian distribution with the mean and standard deviation to train the network based on TID2013. In addition, we adjust the score range of the TID2013 dataset from [0, 9] to [1, 10] to match

the range of the AVA dataset (i.e., [1, 10]). After training the predictors, we use only the mean values of the predicted score distributions to enhance the perceptual quality of the upscaled images.

### 3.5. Discussion

The proposed model extends two existing networks: EUSR [10] as an upscaling model and SRGAN [1] as a discriminator. However, the two networks aim at different objectives: EUSR is for better quantitative quality and SRGAN is for better perceptual quality. Our proposed model combines them to ensure both quantitative and perceptual quality, with two newly proposed components: multi-pass upscaling and qualitative score predictors. In summary, our 4PP-EUSR model achieves the following benefits with the aforementioned components:

- Our model can upscale the input images with considering both the quantitative and perceptual quality. While the base EUSR model tries to make the upscaled images similar to the ground-truth ones, the discriminator reinforces it to focus on fine details. Therefore, our model can achieve better quantitative quality than the other methods concentrating on perceptual quality while keeping the perceptual quality similar to theirs. We will thoroughly investigate this in Section 5.1.
- Thanks to the multi-pass upscaling, the proposed model can learn various upscaling patterns, which will be further discussed in Sections 5.2 and 5.3.
- Employing the qualitative score predictors help our model generate perceptually improved images, since they are trained on the dataset that are obtained directly from human raters. We will discuss their benefits in Sections 5.4 and 5.5.

## 4. Training details

We train our model in three phases: pre-training the EUSR model, building qualitative score predictors, and training the EUSR model in a perceptual manner. Our method is implemented on the TensorFlow framework [33].

### 4.1. Pre-training multi-scale super-resolution model

In our method, we employ 32 and one local residual blocks in the residual module and the upscaling part of the EUSR model, respectively. The EUSR model is first pre-trained with the training set of the DIV2K dataset [23] (i.e., 800 images) using the L1 reconstruction loss as in [10]. For each training step, 16 image patches having a size of  $48 \times 48$  pixels are obtained by randomly cropping the training images. Then, one of the upscaling paths (i.e.,  $\times 2$ ,  $\times 4$ , and  $\times 8$ ) is randomly selected and trained at that step. For instance, when the  $\times 2$  path is selected, the parameters of the path of the model are trained to generate the upscaled images having a size of  $96 \times 96$  pixels. The Adam optimization method [34] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\hat{\epsilon} = 10^{-8}$  is used to update the parameters. A total of 1000,000 training steps are executed with an initial learning rate of  $10^{-4}$  and reducing the learning rate by a half for every 200,000 steps.

<sup>1</sup> <http://www.dpchallenge.com>.

#### 4.2. Training qualitative score predictors

Along with pre-training EUSR, we also train the qualitative score predictors explained in Section 3.4. As the base image classifier, we employ MobileNetV2 [30] pre-trained on the ImageNet dataset [35] with a width multiplier of 1. In the original procedure of training NIMA [28], the input image is rescaled to  $256 \times 256$  pixels without considering the aspect ratio and then randomly cropped to  $224 \times 224$  pixels, which is the input size of VGG19 [13] and Inception-v3 [29]. However, these rescaling and cropping processes are not considered in our case because the MobileNetV2 model does not limit the size of an input image. Instead, we set the input resolution of MobileNetV2 to  $192 \times 192$  pixels, which is the output size of the 4PP-EUSR model for input patches having a size of  $48 \times 48$  pixels. In addition, we do not employ the rescaling step and only employ the cropping step to make the input image have a size of  $192 \times 192$  pixels, because the objective of our score predictors is to evaluate the quality of patches, not the whole given image.

As the loss function for training the qualitative score predictors, we employ the squared Earth mover's distance defined in [36] as

$$E(Q^l, Q^{\tilde{l}}) = \sum_i (F_i(Q^l) - F_i(Q^{\tilde{l}}))^2 \quad (1)$$

where  $I$  and  $\tilde{I}$  are the ground-truth and upscaled images, respectively,  $Q^l$  and  $Q^{\tilde{l}}$  are the probability distributions of the qualitative scores obtained from the predictor for the two images, respectively, and  $F_i(\cdot)$  is the  $i$ th element of the cumulative distribution function of the input. The Adam optimization method [34] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 10^{-7}$  is used to train the parameters.

For the aesthetic score predictor, we use about 5000 images of the AVA dataset [31] for validation and the remaining 250,000 images for training. We first train the new last fully-connected layer for five epochs with a batch size of 128 and a learning rate of  $10^{-3}$ , while freezing all other layers. Then, all the layers are fine-tuned for five epochs with a batch size of 32 and a learning rate of  $10^{-5}$ . For the validation images cropped in the center parts, the predictor achieves an average squared Earth mover's distance of 0.079.

For the subjective score predictor, we consider the first three reference images and their degraded versions in the TID2013 dataset [32] (corresponding to 360 score distributions) for validation and the remaining 22 reference images and their degraded versions (corresponding to 2640 score distributions) for training. Similarly to the aesthetic score predictor, we first train the subjective score predictor with freezing all the layers except the new last fully-connected layer for 100 epochs with a batch size of 128 and a learning rate of  $10^{-3}$ . Then, the whole network is fine-tuned for 100 epochs with a batch size of 32 and a learning rate of  $10^{-5}$ . For the validation images cropped in the center parts, the predictor achieves a Spearman's rank correlation coefficient (SROCC) of 0.780.

#### 4.3. Training super-resolution model

Finally, we fine-tune the pre-trained EUSR model together with the discriminator network using the two trained qualitative score predictors. At each training step, the 4PP-EUSR model outputs three upscaled images by a factor of 4. Then, the discriminator is trained to differentiate the ground-truth and upscaled images based on the sigmoid cross entropy loss as in [1]. After updating parameters of the discriminator, the 4PP-EUSR model is trained with six losses defined as follows.

- **Reconstruction loss ( $l_r$ )**. The reconstruction loss represents the main objective of the super-resolution task: each pixel value of the super-resolved image must be as close as possible to that of

the ground-truth image. In our model, this loss is measured by the pixel-by-pixel L1 loss between the ground-truth and generated images, i.e.,

$$l_r = \frac{1}{W \times H} \sum_w \sum_h |I_{w,h} - \tilde{I}_{w,h}| \quad (2)$$

where  $W$  and  $H$  are the width and height of the images, respectively, and  $I_{w,h}$  and  $\tilde{I}_{w,h}$  are the pixel values at  $(w, h)$  of the ground-truth and upscaled images, respectively.

- **Adversarial loss ( $l_g$ )**. The output of the discriminator network is used to train the super-resolution model towards enhancing perceptual quality, which is denoted as the adversarial loss. It is calculated by the sigmoid cross entropy of the logits obtained from the discriminator for the upscaled images [1]:

$$l_g = -\log(D^{\tilde{l}}) \quad (3)$$

where  $D^{\tilde{l}}$  is the output of the discriminator for the upscaled image  $\tilde{l}$ , which represents the probability that the given image is a real one.

- **Aesthetic score loss ( $l_{as}$ )**. We obtain the aesthetic scores of both the ground-truth and upscaled images from the trained aesthetic score predictor. Then, we define the aesthetic score loss as the weighted difference between the scores, i.e.,

$$l_{as} = \max(0, (S_{a,max} - S_a^{\tilde{l}}) - \alpha_{as}(S_{a,max} - S_a^l)) \quad (4)$$

where  $S_a^l$  and  $S_a^{\tilde{l}}$  are the predicted aesthetic scores of the ground-truth and upscaled images, respectively.  $S_{a,max}$  is the maximum aesthetic score, which is 10 in our case. The term  $\alpha_{as}$  plays a role to control the expected level of aesthetic quality of the upscaled image. For example,  $\alpha_{as} < 1.0$  enforces the model to generate an image that is even perceptually better than the ground-truth image. In our experiments, we set  $\alpha_{as}$  to 0.8.

- **Aesthetic representation loss ( $l_{ar}$ )**. Inspired by Ledig et al. [1], we also define the aesthetic representation loss, which is the L2 loss between the intermediate outputs of the “global average pooling” layer in the aesthetic score predictor for both the ground-truth and upscaled images:

$$l_{ar} = \sum_i (P_{a,i}^l - P_{a,i}^{\tilde{l}})^2 \quad (5)$$

where  $P_{a,i}^l$  and  $P_{a,i}^{\tilde{l}}$  are the  $i$ -th values of the intermediate outputs for the ground-truth and upscaled images, respectively. The length of each intermediate output is 1280 [30].

- **Subjective score loss ( $l_{ss}$ )**. In the same manner as the aesthetic score loss, we calculate the subjective score loss using the trained subjective score predictor, i.e.,

$$l_{ss} = \max(0, (S_{s,max} - S_s^{\tilde{l}}) - \alpha_{ss}(S_{s,max} - S_s^l)) \quad (6)$$

where  $S_s^l$  and  $S_s^{\tilde{l}}$  are the predicted subjective scores of the ground-truth and upscaled images, respectively.  $S_{s,max}$  is the maximum subjective score, which is 10 in our case. Similarly to  $\alpha_{as}$ , the term  $\alpha_{ss}$  controls the contribution of  $S_s^l$ , which is set to 0.8 in our experiments.

- **Subjective representation loss ( $l_{sr}$ )**. In the same manner as the aesthetic representation loss, we calculate the subjective representation loss using the subjective score predictor as

$$l_{sr} = \sum_i (P_{s,i}^l - P_{s,i}^{\tilde{l}})^2 \quad (7)$$

where  $P_{s,i}^l$  and  $P_{s,i}^{\tilde{l}}$  are the  $i$ -th values of the intermediate outputs at the “global average pooling” layer for the ground-truth and upscaled images, respectively.

The losses are calculated for all the three upscaled images and then averaged.

We use the 800 training images of the DIV2K dataset as in the pre-training phase. The Adam optimization method [34] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 10^{-8}$  is used to train both the 4PP-EUSR and discriminator. At every training step, two input image patches are selected, which results in generating six up-scaled images. Thus, the effective batch sizes of the upscaling and discriminative models are six and eight (i.e., two ground-truth and six up-scaled images), respectively. A total of  $4 \times 10^5$  steps are executed with learning rates of  $10^{-5}$  and  $2 \times 10^{-5}$  for the 4PP-EUSR and discriminator, respectively.

## 5. Results

In this section, we report the results of five experiments: comparing the performance of our method and other state-of-the-art super-resolution models, comparing the outputs obtained from different upscaling paths, comparing the performance of our method trained with and without multi-pass upscaling, investigating the roles of loss functions, and comparing the results obtained from different combinations of the loss weights. For the first four experiments, we train our model with the following weighted sum of the six losses defined in Section 4.3:

$$l = 0.05l_r + 0.1l_g + 0.01l_{as} + 0.1l_{ar} + 0.01l_{ss} + 0.1l_{sr} \quad (8)$$

which is empirically determined to ensure high perceptual improvement with minimizing degradation of quantitative performance.

We evaluate the super-resolution performance on the Set5 [37], Set14 [38], and BSD100 [39] datasets. Each dataset contains 4, 14, and 100 images, respectively. We employ five performance metrics that are widely used in the literature, including PSNR, SSIM [9], NIQE [16], a no-reference super-resolution (SR) score proposed by Ma et al. [40], and perceptual index (PI) [11]. PSNR and SSIM are for measuring the quantitative quality, and higher values mean better quality. NIQE, the SR score, and PI are for measuring the perceptual quality, and PI is obtained from the combination of NIQE and the SR score, i.e.,

$$\text{PI}(\tilde{l}) = \frac{1}{2}((10 - \text{SR}(\tilde{l})) + \text{NIQE}(\tilde{l})) \quad (9)$$

where  $\tilde{l}$  is a given up-scaled image, and  $\text{NIQE}(\cdot)$  and  $\text{SR}(\cdot)$  are the measured NIQE value and the SR score, respectively. For NIQE and PI, lower values mean better quality. For the SR score, higher values mean better quality. All quality metrics are calculated on the Y channel of the YCbCr channels converted from the RGB channels with cropping 4 pixels of each border, as in many existing studies [1,10,21]. In addition, we conduct a subjective test to assess the performance of our method in the perspective of real human observers.

### 5.1. Comparison with existing models

We first compare the result images obtained from the  $\times 4$  path of our model with those by the following existing super-resolution models.

- *Bicubic interpolation.* It is a traditional upscaling method, which interpolates pixel values based on values of their adjacent pixels.
- *SRResNet* [1]. This is for single-scale super-resolution, which consists of several residual blocks. Its two variants are considered: The SRResNet-MSE model is trained with the mean-squared loss and the SRResNet-VGG22 model is trained with the Euclidean distance-based loss for the output of the second

*conv3-128* layer of VGG19. Their results are retrieved from the authors' supplementary material<sup>2</sup>.

- *EDSR* [21]. This model also consists of residual blocks similarly to SRResNet, but does not employ batch normalization to improve the performance. In addition, the up-scaled results are obtained by a so-called "geometric self-ensemble" strategy, which obtains eight geometrically transformed versions of the input image via flipping and rotation and blends the model outputs for them. The compared results are obtained from a model trained on the DIV2K dataset, which is provided by the authors.<sup>3</sup>
- *MDSR* [21]. It is an extended version of EDSR, which supports multiple factors of upscaling. We obtain the up-scaled images from the  $\times 4$  path of the MDSR model trained on the DIV2K dataset [23]. The trained model is provided by the authors.<sup>4</sup>
- *EUSR* [10]. This is the base model of 4PP-EUSR, which supports multi-scale super-resolution and consists of optimized residual modules as explained in Section 3.1. We consider the pre-trained EUSR model described in Section 4.1 as a baseline.
- *RCAN* [25]. The RCAN model employs a channel attention mechanism along with the residual-in-residual structure, which contributes to output better super-resolved images than EDSR in terms of PSNR. In addition, RCAN employs the self-ensemble strategy to improve the performance as in the EDSR model. We obtain the output images from the trained model provided by the authors.<sup>5</sup>
- *SRGAN* [1]. The SRGAN model is an extended version of the SRResNet model, where a discriminator network is added to improve the perceptual quality of the up-scaled outputs. We consider three SRGAN models, which use different loss functions to train the discriminator: SRGAN-MSE (the mean-squared loss), SRGAN-VGG22 (the Euclidean distance-based loss for the output of the second *conv3-128* layer of VGG19), and SRGAN-VGG54 (the Euclidean distance-based loss for the output of the fourth *conv3-512* layer of VGG19). The compared results are retrieved from the authors' supplementary material.<sup>6</sup>
- *CX* [27]. This model is based on SRGAN but employs an additional loss function, the contextual loss [41], which measures the cosine distance between the VGG19 features for the ground-truth and up-scaled images. The compared results are retrieved from the authors' website<sup>7</sup>.

**Table 1** compares properties of the baselines and ours, including the number of model parameters, the existence of a multi-scale structure, whether to use the reconstruction loss, whether to employ the discriminator, whether to compare features obtained from well-known image classifiers (e.g., VGG19), and whether to use perceptual scores. First, the EDSR model consists of the largest number of parameters than the other models, while the SRResNet, SRGAN, and CX models have the smallest number of parameters. Our model contains a smaller number of parameters than the MDSR and RCAN models. In terms of the multi-scale structure, MDSR, EUSR, and our model utilize multiple scaling factors, while the other models are based on single-scale super-resolution. Although all the models except SRResNet-VGG22 employ the reconstruction loss, the SRGAN-VGG22 and SRGAN-VGG54 models use it only for pre-training. In addition, SRGANs, CX, and our model employ discriminator networks and use them for adversarial losses. SRResNet-MSE, SRGAN-VGG22, SRGAN-VGG54, and

<sup>2</sup> <https://twitter.com/s/lcue6vldr01ljktdkhhmfvk7vtjhetog>.

<sup>3</sup> [https://cv.snu.ac.kr/research/EDSR/model\\_pytorch.tar](https://cv.snu.ac.kr/research/EDSR/model_pytorch.tar).

<sup>4</sup> [https://cv.snu.ac.kr/research/EDSR/model\\_pytorch.tar](https://cv.snu.ac.kr/research/EDSR/model_pytorch.tar).

<sup>5</sup> <https://github.com/yulunzhang/RCAN>.

<sup>6</sup> <https://twitter.com/s/lcue6vldr01ljktdkhhmfvk7vtjhetog>.

<sup>7</sup> <http://cgm.technion.ac.il/people/Roey/index.html>.

**Table 1**

Properties of the baseline and our models with respect to the number of parameters, multi-scale structure, and loss functions.

Models	# parameters	Multi-scale structure	Using reconstruction loss	Using adversarial loss	Using feature-based loss	Using perceptual loss
SRResNet-MSE	1.5M	No	Yes	No	No	No
SRResNet-VGG22	1.5M	No	No	No	Yes	No
EDSR	43.1M	No	Yes	No	No	No
MDSR	8.0M	Yes	Yes	No	No	No
EUSR	6.3M	Yes	Yes	No	No	No
RCAN	15.6M	No	Yes	No	No	No
SRGAN-MSE	1.5M	No	Yes	Yes	No	No
SRGAN-VGG22	1.5M	No	Yes <sup>a</sup>	Yes	Yes	No
SRGAN-VGG54	1.5M	No	Yes <sup>a</sup>	Yes	Yes	No
CX	1.5M	No	Yes	Yes	Yes	No
<b>4PP-EUSR (ours)</b>	6.3M	Yes	Yes	Yes	Yes	Yes

<sup>a</sup> For pre-training.

**Table 2**

Performance comparison of the baselines and our model evaluated on the Set5 [37], Set14 [38], and BSD100 [39] datasets. The models are sorted by PSNR (dB) in an ascending order.

	PSNR (dB)	SSIM	NIQE	SR score	PI
<b>Set5</b>					
Bicubic	28.418	0.810	8.540	3.770	7.385
CX	29.102	0.830	4.546	7.957	3.295
SRGAN-VGG54	29.410	0.834	4.651	7.940	3.355
SRGAN-VGG22	29.871	0.835	4.919	7.534	3.692
SRResNet-VGG22	30.501	0.869	6.905	6.336	5.285
SRGAN-MSE	30.666	0.859	4.997	7.308	3.844
<b>4PP-EUSR (ours)</b>	31.369	0.870	5.366	6.890	4.238
SRResNet-MSE	32.058	0.892	7.194	5.411	5.891
EUSR	32.352	0.896	7.070	5.173	5.949
MDSR	32.533	0.898	7.111	5.109	6.001
EDSR	32.630	0.899	7.235	5.211	6.012
RCAN	32.713	0.899	7.229	5.277	5.976
<b>Set14</b>					
CX	26.011	0.700	3.460	7.942	2.759
Bicubic	26.091	0.705	7.764	3.661	7.051
SRGAN-VGG54	26.114	0.696	3.875	8.111	2.882
SRGAN-VGG22	26.529	0.712	4.221	7.983	3.119
SRGAN-MSE	27.006	0.719	4.005	7.877	3.064
SRResNet-VGG22	27.272	0.742	7.023	7.093	4.965
<b>4PP-EUSR (ours)</b>	27.969	0.751	4.147	7.457	3.345
SRResNet-MSE	28.590	0.782	6.075	5.648	5.213
EUSR	28.750	0.786	6.168	5.467	5.351
MDSR	28.895	0.789	6.267	5.311	5.478
RCAN	28.951	0.790	6.343	5.451	5.446
EDSR	28.953	0.790	6.305	5.379	5.463
<b>BSD100</b>					
CX	24.581	0.644	3.301	8.801	2.250
SRGAN-VGG54	25.176	0.641	3.407	8.705	2.351
SRGAN-VGG22	25.697	0.660	3.750	8.488	2.631
Bicubic	25.957	0.669	7.712	3.723	6.995
SRGAN-MSE	25.981	0.643	4.032	8.428	2.802
SRResNet-VGG22	26.322	0.694	7.805	7.439	5.183
<b>4PP-EUSR (ours)</b>	26.904	0.701	3.820	7.907	2.956
SRResNet-MSE	27.601	0.737	6.240	5.807	5.217
EUSR	27.674	0.740	6.423	5.808	5.307
MDSR	27.771	0.743	6.538	5.690	5.424
EDSR	27.796	0.744	6.432	5.779	5.326
RCAN	27.821	0.745	6.451	5.868	5.292

CX employ VGG19 as an additional network to use its intermediate outputs as feature-based losses. Our model employs the MobileNetV2-based networks instead of VGG19. Finally, ours estimates the aesthetic and subjective quality scores of the ground-truth and upscaled images for calculating perceptual losses.

Table 2 shows the performance comparison of the baselines and ours evaluated on the three datasets. First of all, the bicubic interpolation introduces a large amount of distortion, which results in low PSNR values, and the upscaled images have poor perceptual quality, according to the high PI values. The models that do not employ a discriminator network (i.e., SRResNet, EDSR, MDSR,

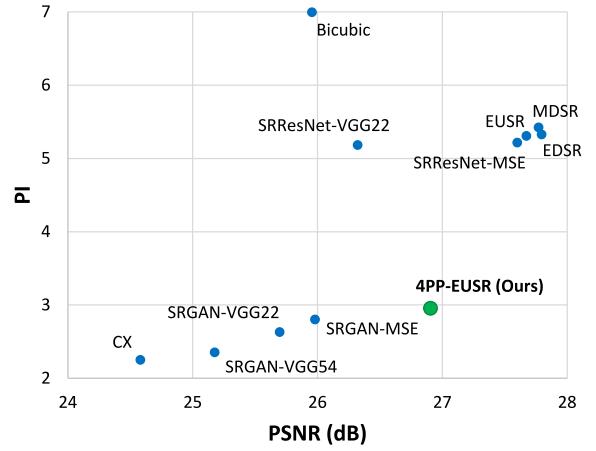
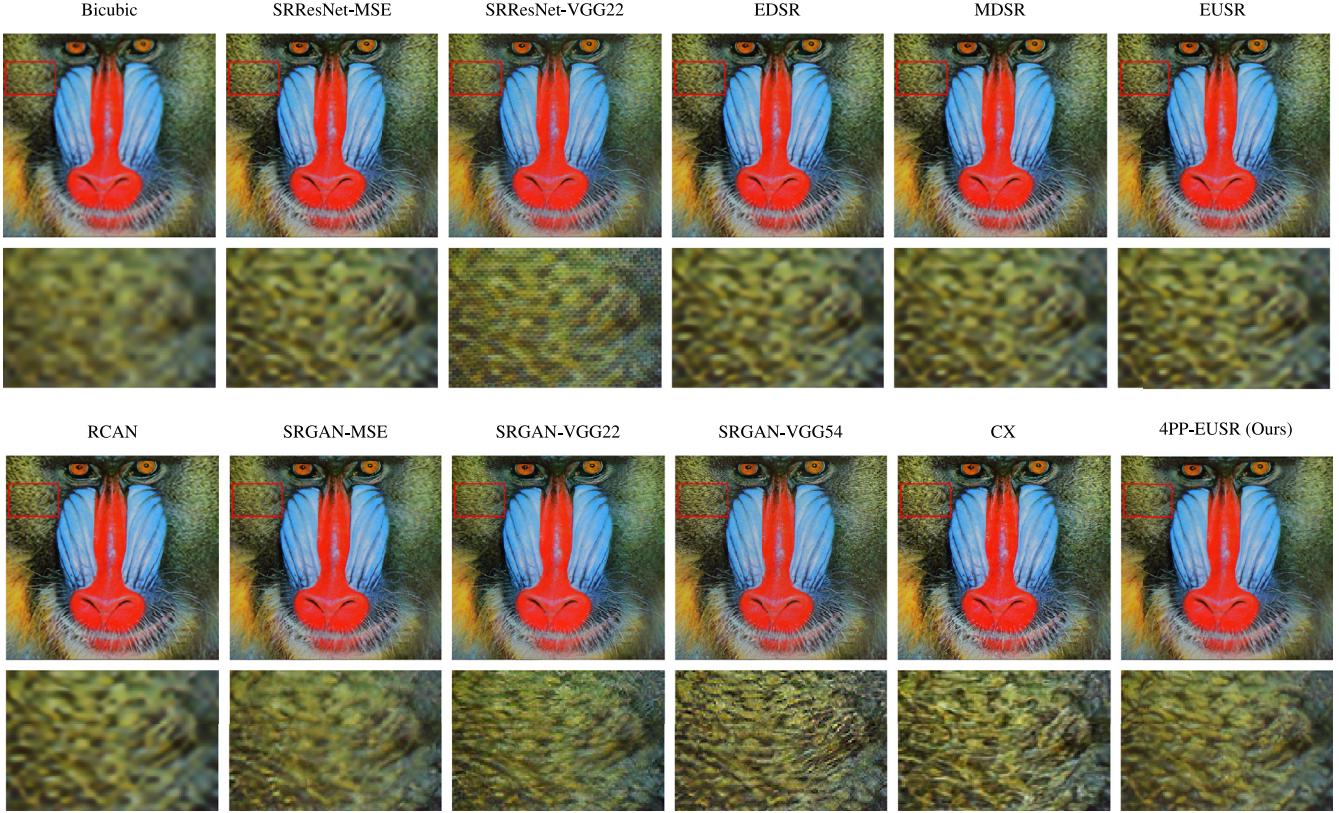


Fig. 6. PSNR and PI values of the baselines and our model for the BSD100 dataset [39].

EUSR, and RCAN) achieve better quantitative quality than the others, showing higher PSNR values, but their perceptual quality is worse except the bicubic interpolation, showing higher PI values. The models considering perceptual quality (i.e., SRGAN and CX) have similar or only slightly higher PSNR values in comparison to the bicubic interpolation, but their perceptual quality is far better than that of the bicubic interpolation, according to the much lower PI values. Our model (i.e., 4PP-EUSR) always records PSNR values higher than those of the other discriminator-based models, which means that ours generates quantitatively better upscaled outputs. At the same time, our model achieves perceptual quality similar to that of SRGAN-MSE in terms of the PI value. For instance, for the BSD100 dataset, the PI values of our model and SRGAN-MSE are 2.956 and 2.802, respectively. This appears more clearly in Fig. 6, which compares the baselines and our model with respect to PSNR and PI values measured for the BSD100 dataset. It confirms that our model achieves proper balances of the quantitative and qualitative quality of the upscaled images.

Fig. 7 shows example images upscaled by different methods. Enlarged images of the regions marked by red rectangles are also shown, where high-frequency textures are expected. First, the bicubic interpolation fails to resolve the textures, producing a highly blurred output. The SRResNet-based, EDSR, MDSR, EUSR, and RCAN models produce richer textures in that region, but still largely blurry. The output of SRResNet-VGG22 shows distinctive textures, which is due to the employment of a different loss function (i.e., differences of VGG19 features). Thanks to the adversarial loss, the other models, including SRGANs, CX, and 4PP-EUSR, generate much better outputs in terms of perceptual quality with sacrificing quantitative quality. Among them, SRGAN-VGG54 and CX recover the most detailed textures, while SRGAN-MSE produces



**Fig. 7.** Images reconstructed by the baselines and our model. The input images are from the Set14 dataset [38].

blurry textures. Our model, 4PP-EUSR, restores the textures more clearly than SRGAN-VGG22 and less distinctly than SRGAN-VGG54. Nevertheless, ours achieves better quantitative quality than all the SRGANs in terms of PSNR in [Table 2](#).

Another comparison shown in [Fig. 8](#) further supports the importance of considering both the quantitative and perceptual quality. Similarly to [Fig. 7](#), the bicubic interpolation shows the worst output than the others, the models employing only the reconstruction loss (i.e., SRResNets, EDSR, MDSR, EUSR, and RCAN) flatten most of the textured areas, and the rest (i.e., SRGANs, CX, and ours) produce outputs having detailed textures. However, the SRGAN and CX models tend to exaggerate the indistinct textures on the ground and airplane regions, introducing sizzling artifacts. For example, the SRGAN-MSE model adds a considerable amount of undesirable noises over the whole image. On the other hand, thanks to the cooperation of the loss functions, our model successfully recovers much of the textures without any prominent artifacts.

In addition, we conduct a subjective test to examine the perceptual performance of the super-resolution methods. We compare the performance of the 12 super-resolution methods for selected ten images in the BSD100 dataset. We employ 15 participants, which meets the required number of participants for subjective tests in the recommendation ITU-R BT.500-13 [42]. As for the evaluation method, we follow the same procedure used in [11]: for a given test image, each participant is asked to rate each of the 120 images on a four-point scale ranging among 1 (definitely fake), 2 (probably fake), 3 (probably real), and 4 (definitely real).

[Fig. 9](#) summarizes the result of the subjective test. It demonstrates that our model outperforms the other methods in terms of the mean opinion score. Our model gets a mean opinion score of 3.07, which means that people regard the output images of ours as “probably real” ones. SRGAN-MSE and SRResNet-VGG22 get the lowest opinion scores among the compared methods. As shown in [Fig. 8](#), it is due to the excessive amount of undesirable artifact

**Table 3**

Performance comparison of the outputs obtained from different three upscaling paths of the 4PP-EUSR model. The results are for the Set5 [37], Set14 [38], and BSD100 [39] datasets.

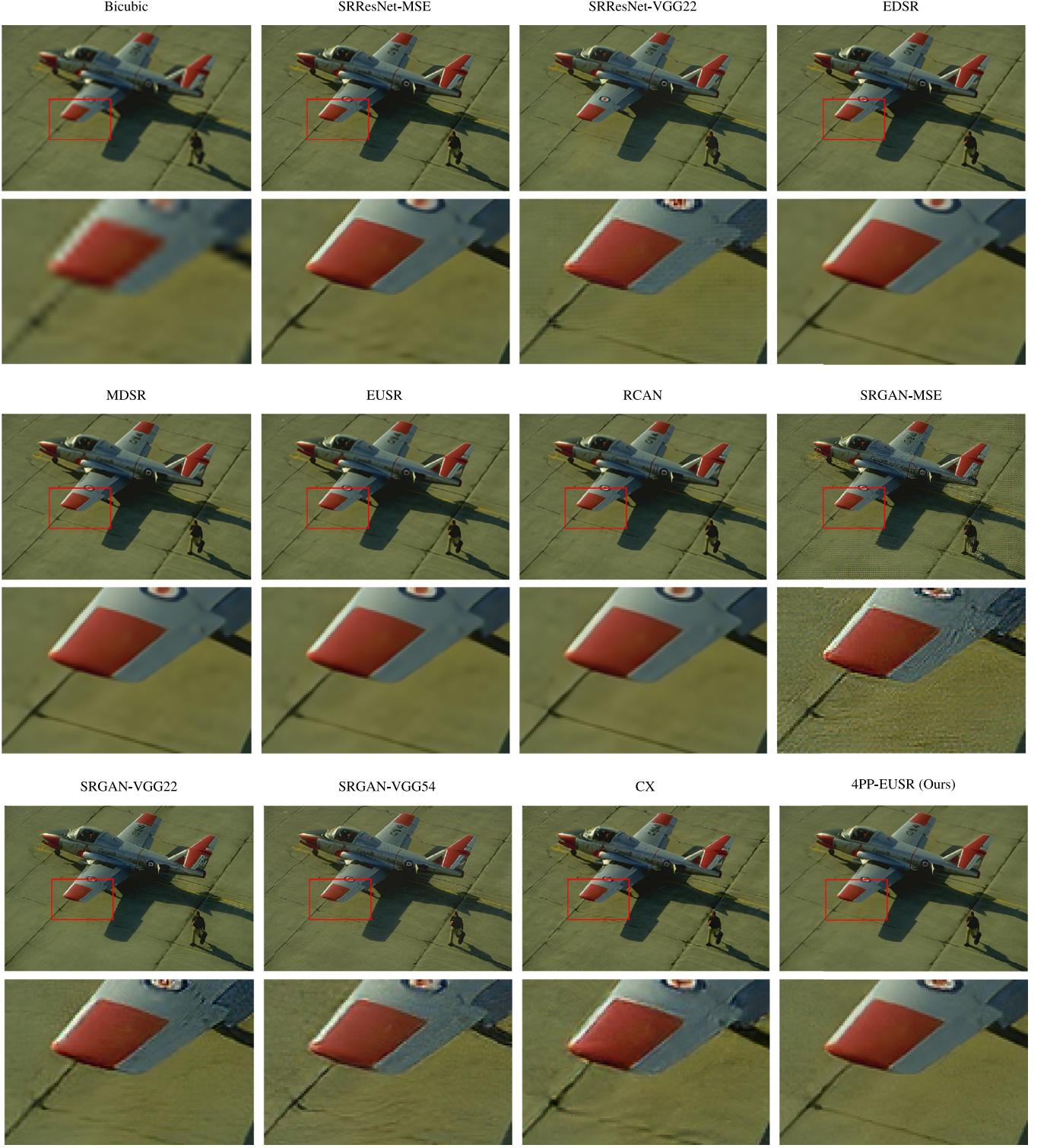
Set5	PSNR (dB)	SSIM	NIQE	SR score	PI
$\times 4$	31.369	0.870	5.366	6.890	4.238
$\times 2 - \times 2$	31.491	0.875	6.500	6.887	4.806
$\times 8 - \text{downscale}$	31.255	0.867	6.044	7.008	4.518
Set14	PSNR (dB)	SSIM	NIQE	SR score	PI
$\times 4$	27.969	0.751	4.147	7.457	3.345
$\times 2 - \times 2$	28.096	0.759	4.858	7.429	3.714
$\times 8 - \text{downscale}$	27.906	0.750	4.631	7.684	3.474
BSD100	PSNR (dB)	SSIM	NIQE	SR score	PI
$\times 4$	26.904	0.701	3.820	7.907	2.956
$\times 2 - \times 2$	27.080	0.710	4.951	7.812	3.570
$\times 8 - \text{downscale}$	26.844	0.699	4.584	8.156	3.214

introduced in the super-resolved images. The result supports that considering both quantitative and perceptual quality in our model is helpful to obtain visually pleasant upscaled images.

## 5.2. Comparing upscaling paths

As described in [Section 3.2](#) and shown in [Fig. 4](#), our model produces three upscaled images by utilizing all the upscaling paths: by passing through the  $\times 4$  path, by passing two times through the  $\times 2$  path, and by passing through the  $\times 8$  path and then downscaling via bicubic interpolation. Here, we compare the results obtained from the different upscaling paths to examine what aspects our model considers to learn.

[Table 3](#) compares the performance of the three upscaling paths of our model. While the PSNR and SSIM values are very similar among the three cases, the  $\times 4$  path shows the best performance



**Fig. 8.** Images reconstructed by the baselines and our model. The input images are from the BSD100 dataset [39].

in terms of the NIQE and PI values. This implies that upscaling using the  $\times 2$  path or  $\times 8$  path is more difficult than the  $\times 4$  path.

Fig. 10 shows an example result showing large differences between the three cases. The appearances of the textures in the enlarged regions are different depending on the upscaling paths, although the overall patterns of the textures follow that of the ground-truth image. First, the output obtained by the two-pass upscaling using the  $\times 2$  path contains grid-like textures. One possible reason is due to the uncertainty in the order of passing: the model

does not know whether the current input image is firstly or secondly inputted between the two passes, thus the two-pass upscaling is not fully optimized. Second, the output obtained from the  $\times 8$  path with downscaling has unexpected white and black pixels, which are similar to the salt-and-pepper noise. It seems that since such noise tends to be removed by downscaling, inclusion of the noise in the output is not necessarily avoided during the training of the  $\times 8$  path. These results show that each upscaling path of our model learns a different strategy for super-resolution and thus

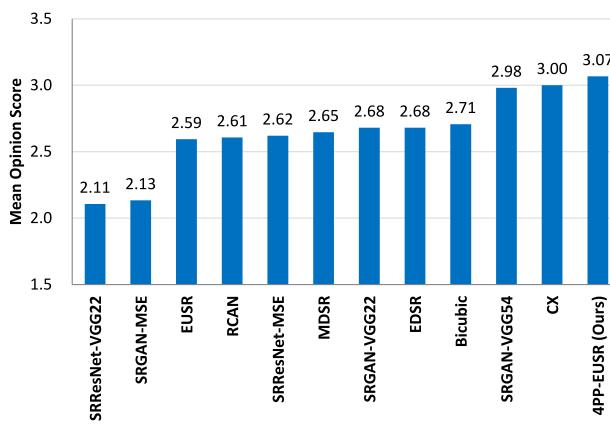


Fig. 9. Subjective test results for 10 images of the BSD100 dataset [39].

Table 4

Performance comparison of the 4PP-EUSR models trained with and without multi-pass upscaling for the Set5 [37], Set14 [38], and BSD100 [39] datasets.

Set5	PSNR (dB)	SSIM	NIQE	SR score	PI
With multi-pass	31.369	0.870	5.366	6.890	4.238
Without multi-pass	31.320	0.869	5.917	6.835	4.541
Set14	PSNR (dB)	SSIM	NIQE	SR score	PI
With multi-pass	27.969	0.751	4.147	7.457	3.345
Without multi-pass	27.699	0.742	4.221	7.594	3.313
BSD100	PSNR (dB)	SSIM	NIQE	SR score	PI
With multi-pass	26.904	0.701	3.820	7.907	2.956
Without multi-pass	26.614	0.688	4.327	8.140	3.093

the model is trained to cope with various types of textures via the shared part of the upscaling paths (i.e., the intermediate residual module shown in Fig. 3).

### 5.3. Effectiveness of multi-pass upscaling

The 4PP-EUSR model employs multi-pass upscaling as aforementioned in Section 3.2. To investigate its effectiveness, we compare the performance of the models trained with and without multi-pass upscaling.

Table 4 shows the performance measures of the models in terms of the PSNR, SSIM, NIQE, SR score, and PI values. It demonstrates that employing multi-pass upscaling is beneficial to enhance both the quantitative and perceptual quality. The model trained with multi-pass upscaling shows larger PSNR and SSIM val-

Table 5

Performance comparison of the 4PP-EUSR models trained by excluding specific loss functions. The models are evaluated on the Set5 [37], Set14 [38], and BSD100 [39] datasets.

	PSNR (dB)	SSIM	NIQE	SR score	PI
With all losses	31.369	0.870	5.366	6.890	4.238
Without $l_r$	29.252	0.834	5.121	8.434	3.344
Without $l_g$	32.145	0.891	6.665	5.687	5.489
Without $l_{as}, l_{ar}$	30.974	0.862	5.503	7.432	4.035
Without $l_{ss}, l_{sr}$	31.389	0.873	5.406	6.807	4.300
Set14	PSNR (dB)	SSIM	NIQE	SR score	PI
With all losses	27.969	0.751	4.147	7.457	3.345
Without $l_r$	26.137	0.705	4.187	8.132	3.028
Without $l_g$	28.589	0.779	5.287	6.153	4.567
Without $l_{as}, l_{ar}$	27.601	0.738	3.976	7.804	3.086
Without $l_{ss}, l_{sr}$	27.853	0.752	4.026	7.571	3.228
BSD100	PSNR (dB)	SSIM	NIQE	SR score	PI
With all losses	26.904	0.701	3.820	7.907	2.956
Without $l_r$	25.142	0.649	4.118	8.773	2.673
Without $l_g$	27.546	0.734	5.362	6.403	4.480
Without $l_{as}, l_{ar}$	26.540	0.684	4.016	8.343	2.837
Without $l_{ss}, l_{sr}$	26.870	0.703	3.780	7.989	2.895

ues and smaller NIQE values for all the three datasets, and smaller PI values for the datasets except Set14. This confirms that the multi-pass upscaling can improve the overall quality of the up-scaled images.

### 5.4. Roles of loss functions

Our model employs multiple types of loss functions as described in Section 4.3. To analyze the role of each loss function, we conduct an experiment where our model is trained with excluding specific loss functions. In detail, we obtain the models trained without  $l_r$ , without  $l_g$ , without  $l_{as}$  and  $l_{ar}$ , and without  $l_{ss}$  and  $l_{sr}$ .

Table 5 shows the PSNR, SSIM, NIQE, SR score, and PI values of the trained models. First, excluding  $l_r$  deteriorates the quantitative quality of the up-scaled images, showing smaller PSNR values, and improves the perceptual quality, showing smaller PI values, in comparison to the model trained with all losses. Excluding  $l_g$  results in the opposite outcomes: it increases the quantitative quality (i.e., larger PSNR values) and decreases the perceptual quality (i.e., larger PI values). Excluding the aesthetic losses (i.e.,  $l_{as}$  and  $l_{ar}$ ) or subjective losses (i.e.,  $l_{ss}$  and  $l_{sr}$ ) also affects to the performance in terms of PSNR.

Fig. 11 shows example output images, where more evident differences of the roles of the loss functions can be observed. First, the image obtained from the model trained without the recon-

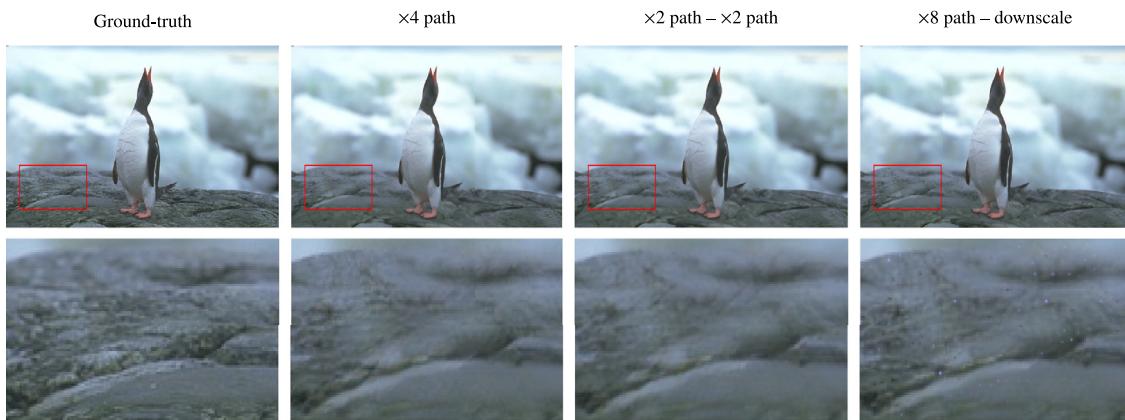
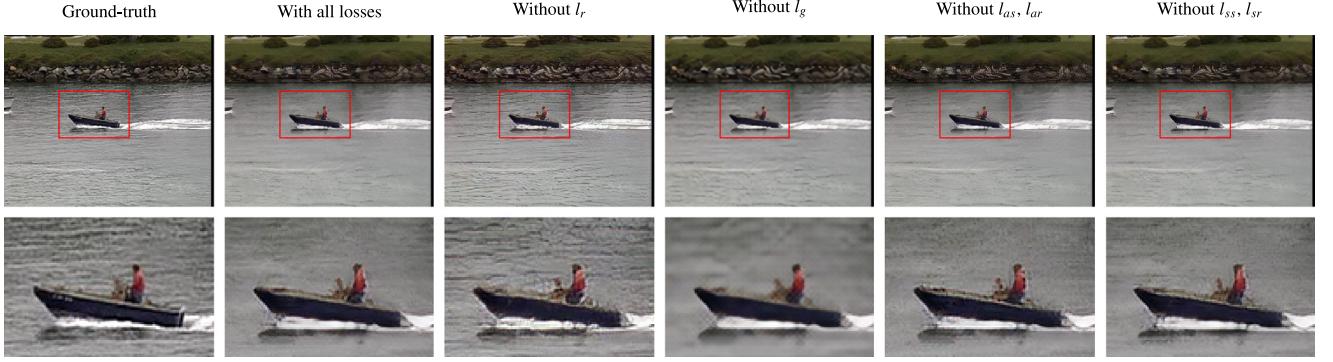


Fig. 10. Images reconstructed by different upscaling paths of our model. The input and ground-truth images are from the BSD100 dataset [39].



**Fig. 11.** Images reconstructed by our models trained with excluding specific loss functions. The input and ground-truth images are from the Set14 dataset [38].

struction loss (i.e.,  $l_r$ ) contains the most distinct textures than the others, but the overall color distribution is slightly different from that of the ground-truth image. On the other hand, the result generated by the model trained without the adversarial loss (i.e.,  $l_g$ ) preserves the overall structure of the ground-truth image, while its details are more blurry than those of the others. The output of the model trained without the subjective loss functions contains more lattice-like textures than that of the model trained without the aesthetic loss functions. This implies that the aesthetic losses contribute to the restoration of highly structured textures, while the subjective losses are helpful to construct dispersed high-frequency textures. Finally, the image obtained by training with all the proposed loss functions is the most reliable and natural.

### 5.5. Comparing different loss weights

Finally, we train our model with different weights of the loss functions. Specifically, we alter the weight of the reconstruction loss in (8) as

$$l = \alpha_r l_r + 0.1 l_g + \alpha_p (0.01 l_{as} + 0.1 l_{ar} + 0.01 l_{ss} + 0.1 l_{sr}) \quad (10)$$

with  $\alpha_r \in \{0.5, 0.05, 0.005\}$  and  $\alpha_p \in \{0, 1\}$ . We can expect that a larger  $\alpha_r$  value leads the model to be trained towards producing outputs having better quantitative quality. The term  $\alpha_p$  determines whether to use the score predictors or not.

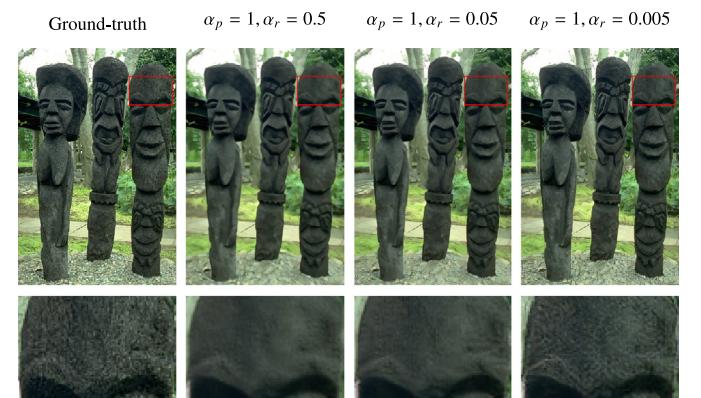
**Table 6** presents the performance of our model trained with different weight values. As expected, decreasing the level of contribution of the reconstruction loss with a smaller  $\alpha_r$  results in lower PSNR values. On the other hand, the PI values are also decreased, which indicates improved qualitative quality. These observations emerge as the visual differences of the upscaled images shown in **Fig. 12**. When we examine the enlarged regions where high-frequency textures are expected, a decreased  $\alpha_r$  value affects the clearness of the output images, due to relatively larger contributions of the adversarial and perceptual losses. These confirm that there is a tradeoff between quantitative and perceptual quality as mentioned in [14], and our model has a capability to deal with the priorities of these quality measures by adjusting the weights of the loss functions.

In addition, the result shows that employing the aesthetic and subjective score predictors (i.e., with  $\alpha_p = 1$ ) is helpful to improve the perceptual quality of the upscaled images, which can be observed from increased SR scores and decreased NIQE and PI values in comparison to the cases without the predictors in **Table 6**. This shows that the score predictors are highly effective when both are employed and play a significant role in improving the perceptual quality. By employing the score predictors along with the other loss functions and the multi-pass upscaling, our model can output the super-resolved images with better quality than the other models, as shown in **Section 5.1**.

**Table 6**

Performance comparison of our models trained with different combinations of the loss weights. The models are evaluated on the Set5 [37], Set14 [38], and BSD100 [39] datasets.

Set5	PSNR (dB)	SSIM	NIQE	SR score	PI
$\alpha_p = 0, \alpha_r = 0.5$	31.891	0.881	6.386	5.637	5.375
$\alpha_p = 0, \alpha_r = 0.05$	31.748	0.880	5.739	6.073	4.833
$\alpha_p = 0, \alpha_r = 0.005$	30.504	0.854	5.730	7.633	4.048
$\alpha_p = 1, \alpha_r = 0.5$	31.839	0.881	6.242	5.956	5.143
$\alpha_p = 1, \alpha_r = 0.05$	31.369	0.870	5.366	6.890	4.238
$\alpha_p = 1, \alpha_r = 0.005$	30.753	0.857	5.234	7.685	3.775
Set14	PSNR (dB)	SSIM	NIQE	SR score	PI
$\alpha_p = 0, \alpha_r = 0.5$	28.348	0.765	4.597	6.852	3.872
$\alpha_p = 0, \alpha_r = 0.05$	28.218	0.764	4.154	7.099	3.527
$\alpha_p = 0, \alpha_r = 0.005$	26.864	0.715	4.644	7.897	3.374
$\alpha_p = 1, \alpha_r = 0.5$	28.316	0.763	4.766	6.843	3.961
$\alpha_p = 1, \alpha_r = 0.05$	27.969	0.751	4.147	7.457	3.345
$\alpha_p = 1, \alpha_r = 0.005$	27.020	0.726	4.017	7.970	3.023
BSD100	PSNR (dB)	SSIM	NIQE	SR score	PI
$\alpha_p = 0, \alpha_r = 0.5$	27.332	0.717	4.633	6.932	3.850
$\alpha_p = 0, \alpha_r = 0.05$	27.162	0.715	3.987	7.478	3.254
$\alpha_p = 0, \alpha_r = 0.005$	25.833	0.659	5.374	8.548	3.413
$\alpha_p = 1, \alpha_r = 0.5$	27.271	0.714	4.498	7.042	3.728
$\alpha_p = 1, \alpha_r = 0.05$	26.904	0.701	3.820	7.907	2.956
$\alpha_p = 1, \alpha_r = 0.005$	26.176	0.678	3.867	8.552	2.657



**Fig. 12.** Images reconstructed by our models trained with different combinations of the loss weights. The input and ground-truth images are from the BSD100 dataset [39].

## 6. Conclusion

In this paper, we proposed a perceptually improved super-resolution method, which employs multi-pass image restoration via a multi-scale super-resolution model and trains the model with a discriminator network and two qualitative score predictors. The

results showed that our model successfully recovers the original textures in a perceptual manner while preventing quantitative quality degradation.

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the “ICT Consilience Creative Program” (IITP-2019-2017-0-01015) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). In addition, this work was also supported by the IITP grant funded by the Korea government (MSIT) (R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding).

### References

- [1] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A.P. 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.
- [2] S. Yang, Z. Liu, M. Wang, F. Sun, L. Jiao, Multitask dictionary learning and sparse representation based single-image super-resolution reconstruction, Neurocomputing 74 (17) (2011) 3193–3203.
- [3] X. Li, H. He, Z. Yin, F. Chen, J. Cheng, Single image super-resolution via subspace projection and neighbor embedding, Neurocomputing 139 (2014) 310–320.
- [4] J. Salvador, E. Perez-Pellitero, Naive Bayes super-resolution forest, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 325–333.
- [5] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: a review, Neurocomputing 187 (2016) 27–48.
- [6] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
- [7] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, Q. Liao, Deep learning for single image super-resolution: a brief review, IEEE Trans. Multimedia (2019) (early access).
- [8] C. Dong, C.C. Loy, K. He, X. Tang, Learning a deep convolutional network for image super-resolution, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 184–199.
- [9] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
- [10] J.-H. Kim, J.-S. Lee, Deep residual network with enhanced upscaling module for super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 913–921.
- [11] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, L. Zelnik-Manor, The 2018 PIRM challenge on perceptual image super-resolution, in: Proceedings of the European Conference on Computer Vision Workshops, 2018, pp. 334–355.
- [12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
- [13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Proceedings of the International Conference on Learning Representations, 2015.
- [14] Y. Blau, T. Michaeli, The perception-distortion tradeoff, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6228–6237.
- [15] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.
- [16] A. Mittal, R. Soundararajan, A.C. Bovik, Making a “completely blind” image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2013) 209–212.
- [17] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.
- [18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [19] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the International Conference on Machine Learning, 2015, pp. 448–456.
- [20] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
- [21] B. Lim, S. Son, H. Kim, S. Nah, K.M. 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.
- [22] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, K.M. Lee, et al., NTIRE 2017 challenge on single image super-resolution: Methods and results, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1110–1121.
- [23] R. Timofte, S. Gu, J. Wu, L. Van Gool, L. Zhang, M.-H. Yang, et al., NTIRE 2018 challenge on single image super-resolution: Methods and results, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 965–976.
- [24] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, 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.
- [25] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 286–301.
- [26] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 694–711.
- [27] R. Mechrez, I. Talmi, F. Shama, L. Zelnik-Manor, Maintaining natural image statistics with the contextual loss, Proceedings of the Asian Conference on Computer Vision, 2018, pp. 427–443.
- [28] H. Talebi, P. Milanfar, NIMA: Neural image assessment, IEEE Trans. Image Process. 27 (8) (2018) 3998–4011.
- [29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
- [30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
- [31] N. Murray, L. Marchesotti, F. Perronnin, AVA: A large-scale database for aesthetic visual analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2408–2415.
- [32] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al., Image database TID2013: peculiarities, results and perspectives, Signal Process. Image Commun. 30 (2015) 57–77.
- [33] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., TensorFlow: a system for large-scale machine learning, in: Proceedings of the USENIX Symposium on Operating Systems Design and Implementation, 2016, pp. 265–283.
- [34] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Proceedings of the International Conference on Learning Representations, 2015, pp. 1–13.
- [35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
- [36] L. Hou, C.-P. Yu, D. Samaras, Squared earth mover’s distance-based loss for training deep neural networks, arXiv:1611.05916 (2016) 1–9.
- [37] M. Bevilacqua, A. Roumy, C. Guillemot, M.L. Alberi-Morel, Low-complexity single-image super-resolution based on nonnegative neighbor embedding, in: Proceedings of the British Machine Vision Conference, 2012, pp. 1–10.
- [38] R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: Proceedings of the International Conference on Curves and Surfaces, 2010, pp. 711–730.
- [39] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proceedings of the IEEE International Conference on Computer Vision, 2001, pp. 416–423.
- [40] C. Ma, C.-Y. Yang, X. Yang, M.-H. Yang, Learning a no-reference quality metric for single-image super-resolution, Comput. Visi. Image Underst. 158 (2017) 1–16.
- [41] R. Mechrez, I. Talmi, L. Zelnik-Manor, The contextual loss for image transformation with non-aligned data, Proceedings of the European Conference on Computer Vision, 2018, pp. 768–783.
- [42] ITU, Recommendation ITU-R BT.500-13: Methodology for the subjective assessment of the quality of television pictures, 2012.



**Jun-Ho Choi** received the B.S. degree from the School of Integrated Technology, Yonsei University, Korea, in 2014, where he is currently working toward the Ph.D. degree. His research interests include photo quality assessment, aesthetic multimedia editing, and multimodal activity recognition.



**Jun-Hyuk Kim** received the B.S. degree in the School of Integrated Technology from Yonsei University, Korea, in 2015, where he is currently working toward the Ph.D. degree. His research interests include multimedia signal processing and machine learning.



**Jong-Seok Lee** received the Ph.D. degree in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Korea, in 2006. From 2008 to 2011, he was a Research Scientist with the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland. He is currently an Associate Professor with the School of Integrated Technology, Yonsei University, Korea. He has authored or coauthored more than 100 publications. His research interests include machine learning and multimedia signal processing. He serves as an Editor for IEEE Communications Magazine and Signal Processing: Image Communication.



**Manri Cheon** received the B.S. degree in electrical and electronic engineering and Ph.D. degree in engineering from Yonsei University, Korea, in 2012 and 2018, respectively. He is working as a postdoctoral researcher at the Multimedia Computing and Machine Learning (MCML) Group in Yonsei University, Korea. His current research interests include quality of experience (QoE)-based multimedia signal processing, perceptual quality assessment for image and video, and deep learning for multimedia signal processing.