

Received 3 December 2023, accepted 29 December 2023, date of publication 1 January 2024,
date of current version 8 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3349023

RESEARCH ARTICLE

Semantic Super-Resolution via Self-Distillation and Adversarial Learning

HANHOON PARK 

Division of Electronics and Communications Engineering, Pukyong National University, Busan 48513, Republic of Korea
Department of Artificial Intelligence Convergence, Graduate School, Pukyong National University, Busan 48513, Republic of Korea
e-mail: hanhoon.park@pknu.ac.kr


This work was supported by the National Research Foundation of Korea (NRF) Grant by the Korean Government through the MSIT under Grant 2021R1F1A1045749.

ABSTRACT Semantic super-resolution (SR) is an approach that improves the SR performance by leveraging semantic information about the scene. This study develops a novel semantic SR method that is based on the generative adversarial network (GAN) framework and self-distillation. A discriminator is adversarially trained along with a generator to extract semantic features from images and distinguish semantic differences between images. To train the generator, an additional adversarial loss is computed from the discriminator's outputs of SR images belonging to the same category and minimized via self-distillation. This guides the generator to learn implicit category-specific semantic priors. We conducted experiments for SR of text and face images using the Enhanced Deep Super-Resolution (EDSR) generator and the SRGAN discriminator. Experimental results showed that our method can contribute to improving both the quantitative and qualitative quality of SR images. Although the improvement varied depending on image category and dataset, the peak signal-to-noise ratio (PSNR) value increased by up to 0.87 dB and the perceptual index (PI) decreased by up to 0.17 by using our method. Our method outperformed existing semantic SR methods.

INDEX TERMS Image super-resolution, semantic super-resolution, self-distillation, adversarial learning, text images, face images, EDSR, SRGAN.

I. INTRODUCTION

Super-resolution (SR) is the process of up-sampling a low-resolution (LR) image to recover the underlying high-resolution (HR) image, which has been used in various applications such as surveillance, forensics, microscopy, and remote sensing [1]. Inspired by the great success of convolutional neural network (CNN) in various computer vision approaches [2], recent SR studies have attempted to use CNNs and have produced visually pleasing SR images with low pixel errors [3]. However, CNNs suffer from the well-known over-fitting problem, which reduces the CNNs' ability to generalize unseen data. Therefore, various approaches to enhance generalization have been introduced, and applied to SR studies [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang .

For the purpose of generalization, most CNN-based SR methods usually avoid using only certain categories of images for training because the biased training will cause the over-fitting problem. However, the greater the number of image categories used in the training process, the lower the SR performance for each category. Therefore, an SR method that is free of over-fitting while maximizing the SR performance for each category is required. To this regard, semantic SR methods have attempted to incorporate contextual or semantic information of images belonging to the same category into the SR process, resulting in more accurate SR results without over-fitting. As the most recent study, Park [5] introduced a semantic loss to measure the semantic difference between text images and proposed a semantic SR method that minimizes the semantic loss via self-distillation. The method allowed existing SR models to be trained to produce better text SR images and the Enhanced Deep Super-Resolution (EDSR) model [6] trained

using the method outperformed the generative adversarial network (GAN)-based semantic SR model. However, the method has been validated only for SR of text images. Also, the semantic loss required the pretrained VGG network [7] to extract semantic features from images, and the SR model was trained to directly minimize the semantic loss along with a pixel loss. This caused that the performance of the SR model depended on the feature extraction capability of the VGG network, and the incompleteness of the VGG network as a semantic feature extractor misled the trained SR model to produce degraded SR images. Therefore, this study aims improving the Park's method and proposes a novel semantic SR method. Basically, the proposed method is based on self-distillation as in the Park's method, but does not require any pretrained networks to extract semantic features from images. Instead, we introduce a classifier (= discriminator) that extracts semantic features from images and recognizes semantic differences between them. The classifier is adversarially trained along with an SR generator (= one of existing SR models) to extract semantic features from SR images and convey the semantic information to the generator. In the process of training the generator, we add an adversarial loss that is computed from the classifier output and minimize the loss via self-distillation; thus, unlike the Park's method, the semantic difference between SR images belonging to the same category is indirectly minimized. In addition, our method is validated on text and face image datasets.

The primary contributions of this study, which focuses on developing an effective semantic SR method is as follows:

- We propose a novel semantic SR method that is based on self-distillation and adversarial learning.
- We propose to train additional classifier to extract good semantic features from images and recognize semantically similar images.
- Our method allows existing SR models to be generalized to better super-resolve specific categories' images without over-fitting.
- The performance of our method is validated on different text and face image datasets in various aspects. The experiments demonstrate that our method outperforms other semantic SR methods.

II. RELATED WORK

A. SELF-DISTILLATION

Knowledge distillation is the process that helps train student networks by transferring extra supervised information distilled from the pretrained teacher network [8]. It has been commonly used for network compression. Self-distillation is one of the knowledge distillation techniques, but it focuses on efficiently optimizing a network from the consistent distributions of data representations without the assistance of teacher networks [9]. Therefore, it was successfully used as a regularization technique for matching the predictive distribution of the network between different samples of the same label [10]. Also, it was used for matching semantic

features between different images of the same category and guiding the network to learn the semantic information [5]. We also use self-distillation for the semantic learning in this study.

B. CNN-BASED SR

SR has witnessed great strides with the development of deep learning and CNNs.

The Super-Resolution Convolutional Neural Network (SRCNN) [11] is considered to be the pioneering work in using CNNs for the task of SR. It only consists of three layers and requires the LR image to be up-sampled using bicubic interpolation prior to being processed by the network, but it has been shown to outperform the traditional SR methods that do not use CNNs.

The Residual Network (ResNet) [12], designed to ease the training of networks with a number of layers by adding skip/shortcut connections, was applied to the SR domain to create SRResNet, where it was also used as the generator network of a GAN-based network termed SRGAN [13]. EDSR [6] was also based on ResNet, and removed batch normalization layers to disable restriction of the feature values and reduce memory usage during training, allowing more layers and filters to be used. The ResNet-based deep networks showed significantly improved SR performance.

The Residual Channel Attention Network (RCAN) [14] is a much deeper network, comprising of residual groups that each contains a number of residual channel attention blocks. A long skip connection was used with the residual groups, enabling propagation of information from the early stages to the latter stages of the network, whereas a short skip connection was used inside the residual blocks, serving to propagate information at finer levels. Thus, this residual-in-residual architecture enabled to train very deep CNNs (more than 400 layers) while effectively conveying the low-frequency information across layers.

Most CNN-based SR methods have tried to reduce the pixel error of SR images and yield high peak signal-to-noise ratio (PSNR) values. However, PSNR is known to be poorly correlated with human visual perception [15]. Focused more on the perceptual quality of SR images, SRGAN [13] used GAN with the perceptual loss computed using the pretrained VGG network and produced visually more convincing images.

Enhanced SRGAN (ESRGAN) [16] improved SRGAN by modifying the network architecture and the loss function. It removed batch normalization layers (similarly to EDSR) and used the residual-in-residual dense block as the basic building block of the network to enable higher network capacity and facilitate training. Also, ESRGAN used the relativistic average discriminator, which predicts the probability that an image is relatively more realistic than the other. Finally, to compute the perceptual loss, ESRGAN used the VGG features before activation layers to reduce the sparsity of features and better supervise brightness consistency

and texture recovery. Benefiting from these modifications, ESRGAN produced SR images with sharper edges and better textures, while reducing undesirable artifacts.

After ESRGAN, a number of ESRGAN variants have been proposed [17] and shown to produce SR images with more realistic and natural textures than ESRGAN.

Most recently, attempts to combine CNNs with the Vision Transformer (ViT) [18] have been reported in the SR domain as well. Some studies [19], [20] used a 3×3 convolutional layer to extract shallow features before applying the ViT. This provided a simple way to map the input image space to a higher dimensional feature space and led to more stable optimization and better SR results.

The Efficient Super-Resolution Transformer (ESRT) [21] used CNNs more actively, to address the heavy computational cost and high GPU memory occupation of the ViT. ESRT is composed of the Lightweight Transformer Backbone (LTB) and Lightweight CNN Backbone (LCB). LTB is responsible for capturing long-term spatial dependencies across local regions within an image and LCB is responsible for extracting deep features while dynamically adjusting feature sizes to maintain low computational cost. It was shown that ESRT can achieve a very good balance between performance and computational complexity. As another example of strategically combining CNNs with Transformers, the Transformer-CNN Feature Distillation Network (TCFDN) [22] is a hybrid network of Transformer and CNN with cascaded feature distillation blocks for efficient SR. In each feature distillation block, 1×1 convolutional layers are responsible for distilling features and reducing channels with few parameters, Transformer layers are responsible for attending to spatial context and gradually refining features to attain more discriminate information. TCFDN was able to extract refined multi-level features with better representation ability while remaining lightweight. Other efficient hybrid networks can be found in [23].

In this study, we use SRGAN whose generator is replaced by EDSR as the baseline network for the convenience of implementation. However, our method is applicable to other SR models with minor modifications.

C. SEMANTIC SR

Semantic SR is an approach that generates more accurate SR results by incorporating contextual or semantic information about the scene into the SR process.

In traditional exemplar-based SR, Sun et al. [24] realized context-constrained face image SR by building a training set of texturally similar HR/LR image segment pairs. Timofte et al. [25] investigated the role of semantic priors on SR by training specialized models separately for each semantic category, and showed that semantic information can help the SR process enhance local image details.

In CNN-based SR, Xu et al. [26] first attempted to use semantic priors. They proposed using a multi-class GAN with multiple discriminators, where multiple discriminators help

the generator learn category-specific semantic information of images of different categories (e.g., text and face images). To assume different semantic priors co-existing in an image, Wang et al. [27] obtained semantic priors at the pixel level using a pretrained semantic segmentation network. Then, they proposed a spatial feature transform layer to efficiently incorporate the semantic information into the SR process. Frizza et al. [28] proposed a similar approach based on a typical GAN formulation and a pretrained semantic segmentation network. They modified the network architecture of ESRGAN using the findings of previous CNN studies and trained the generator by adding a semantic similarity loss, which represents the difference between the semantic masks of HR and SR images obtained using the segmentation network. Chen et al. [29] proposed a blind SR method using a semantic-aware texture prior obtained by applying mini-batch K-means to feature vectors of HR images. Recently, to leverage semantic priors of text images, Park [5] proposed minimizing the semantic difference between text SR images using self-distillation. To this end, he proposed a semantic loss that represents semantic differences between images and is computed using the pretrained VGG network. He showed that the learned semantic information can help SR networks produce better text SR images and his method outperformed Xu et al.'s one.

Our method is similar to Park's method, but improves its performance by reformulating it in a GAN framework. Our method is the same as Xu et al.'s method in that it is a GAN-based framework. However, our method does not require multiple discriminators, and it minimizes semantic differences between images belonging to the same category in the SR process by introducing a semantic loss, thereby being able to be more effective for semantic SR.

D. SUMMARY

Various structures and types of SR networks have been proposed, and SR performance has been greatly improved. However, SR performance can be improved more effectively by using semantic information of the image, and it has been shown that using the GAN framework is common and excellent in terms of performance. Therefore, in this study, we also use the GAN framework and utilize self-distillation to effectively learn semantic information without significantly changing the existing network structure. This methodology has not been attempted before, and may have superior performance compared to methods that do not use the GAN framework or require network structure changes.

III. PROPOSED SEMANTIC SR METHOD

Inspired by the Park's method [5], the semantic priors are incorporated into the SR process via self-distillation. That is, we distill the category-specific semantic information from SR images of the same category (e.g., "text" or "face") during training an SR generator. To this end, we propose a semantic loss that enforces SR images belonging to the same category to be semantically or contextually similar. However,

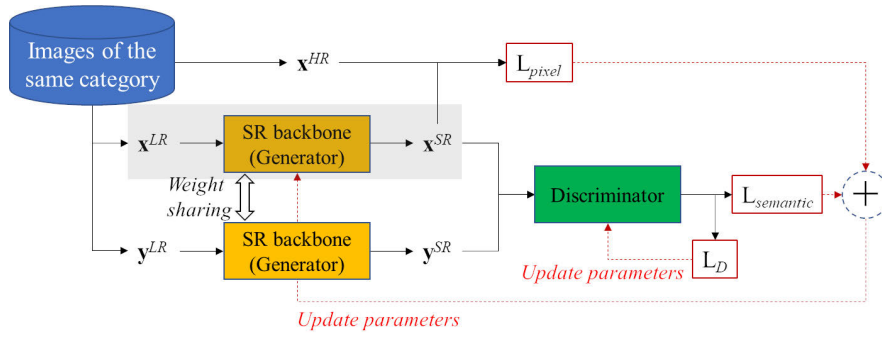


FIGURE 1. Process flow of the proposed method.

the semantic loss is not directly computed using the pretrained VGG network [7]. Instead, we train a discriminator to distinguish semantic differences between SR images belonging to the same category (see Fig. 1). Then, the semantic loss is indirectly computed from the discriminator’s output, aiming to fool the discriminator. In other words, the generator is trained so that its SR images belonging to the same category are semantically similar, by minimizing the semantic loss via adversarial learning. This is because the pretrained VGG network (used in the Park’s method) cannot adequately mine semantic features of images and direct matching of incomplete semantic features may cause the SR process to proceed in the wrong way.

To distinguish semantic differences between SR images belonging to the same category, the discriminator will be trained to extract features, which are most discriminative for images of a specific category. Therefore, using the trained discriminator (instead of the pretrained VGG network) enables to extract better semantic features from images and helps generate more realistic SR images.

As shown in Fig. 1, we use one of existing CNN-based SR models as the backbone network, which plays a role of generating SR images. To train the generator, we randomly sample an LR-HR image pair (x^{LR}, x^{HR}) and another LR-HR image pair (y^{LR}, y^{HR}) belonging to the same category, from the training dataset. Then, we minimize the loss:

$$\mathcal{L}_G = \mathcal{L}_{pixel}(x^{HR}, x^{SR}) + \kappa \mathcal{L}_{semantic}(x^{SR}, y^{SR}), \quad (1)$$

where x^{SR} and y^{SR} represent the outputs (SR images) of the generator given the input x^{LR} and y^{LR} , respectively. κ is a weighting factor and set to 0.001 in our experiments. \mathcal{L}_{pixel} is the pixel loss and defined as:

$$\mathcal{L}_{pixel}(x, y) = \frac{1}{WH} \sum_{i=1}^W \sum_{j=1}^H \|x(i, j) - y(i, j)\|_1, \quad (2)$$

where W and H are the image dimensions. $\mathcal{L}_{semantic}$ is the semantic loss and defined as:

$$\mathcal{L}_{semantic}(x, y) = BCE(D(x), 0) + BCE(D(y), 1), \quad (3)$$

where $D(x)$ denotes the discriminator’s response to input x and $BCE()$ represents the binary cross entropy function.

The discriminator can also be one of existing CNN-based binary classifiers. To train the discriminator, we minimize the loss:

$$\mathcal{L}_D = BCE(D(x^{SR}), 1) + BCE(D(y^{SR}), 0). \quad (4)$$

As a result, the discriminator is trained to distinguish semantic differences between x^{SR} and y^{SR} , whereas the generator is trained to generate x^{SR} semantically similar to y^{SR} by fooling the discriminator.

The training procedure is summarized in Algorithm 1. Only the shaded parts in Fig. 1 are active during inference, so there is no computational overhead for semantic SR.

Algorithm 1 Semantic SR via self-distillation and adversarial learning

Initialize network parameters of the SR backbone network and the discriminator, step $t = 1$.

do

Sample a batch (x^{LR}, x^{HR}) from the training dataset. Sample another batch y^{LR} randomly, which belongs to the same category from the training dataset. Get the SR images of x^{LR} and y^{LR} by feeding them into the SR backbone network. Feed x^{SR} and y^{SR} into the discriminator. Update parameters of the SR backbone network by minimizing the loss function in Eq. 1. Update parameters of the discriminator by minimizing the loss function in Eq. 4.

$t \leftarrow t + 1$.

while $t < T$;

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. EXPERIMENTAL SETUP

The main goal of this study is to verify that our method enables existing SR models to learn semantic priors, thereby improving their performance for certain categories of images. We use EDSR [6] as a baseline model.¹ Therefore, we con-

¹EDSR has been widely used as a baseline model in recent SR methods due to its high performance and effectiveness [30], [31].

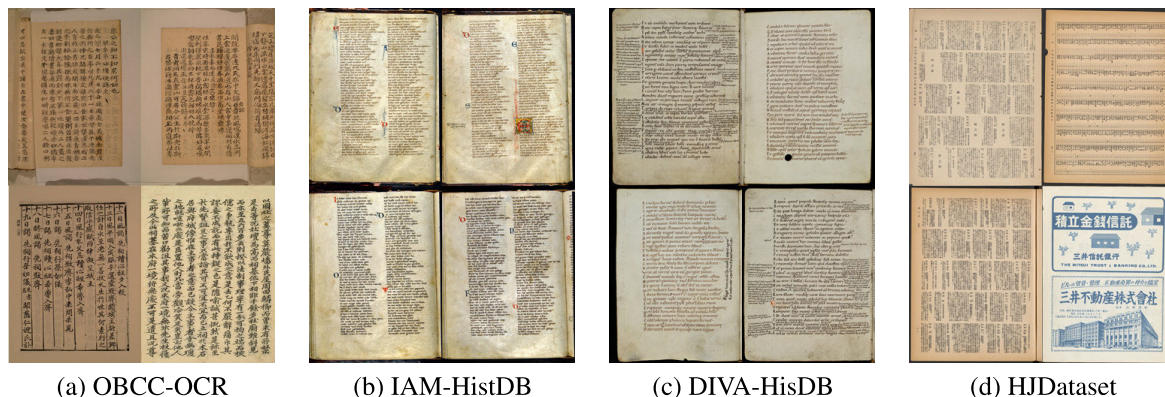


FIGURE 2. A part of images from text image datasets used in our experiments.

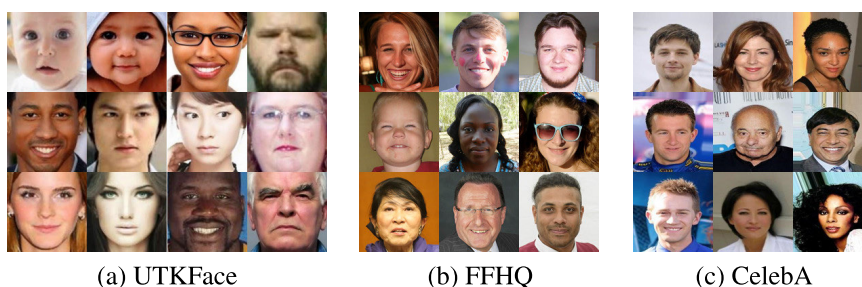


FIGURE 3. A part of images from face image datasets used in our experiments.

TABLE 1. Quantitative image quality evaluation of text SR images produced by different SR methods. The values to the left and right of ‘/’ represent the PSNR and SSIM, respectively.

	OBCC-OCR	IAM-HistDB	DIVA-HisDB	HJDataset
Vanilla EDSR	32.1724 / 0.9372	24.4712 / 0.6715	35.6505 / 0.9179	24.1926 / 0.7678
Xu et al.’s [26]	32.5619 / 0.9383	24.5915 / 0.6754	36.0230 / 0.9260	24.2622 / 0.7715
Park’s [5]	32.2118 / 0.9371	24.6318 / 0.6782	36.2839 / 0.9582	24.2655 / 0.7747
Proposed	32.4196 / 0.9385	24.6436 / 0.6797	36.5206 / 0.9787	24.2970 / 0.7782

TABLE 2. Quantitative image quality evaluation of face SR images produced by different SR methods. The values to the left and right of ‘/’ represent the PSNR and SSIM, respectively.

	UTKFace	FFHQ	CelebA
Vanilla EDSR	36.8009 / 0.9338	34.5944 / 0.9030	30.2996 / 0.8753
Xu et al.’s [26]	36.8506 / 0.9339	34.6122 / 0.9032	30.3785 / 0.8761
Park’s [5]	36.4681 / 0.9307	34.4951 / 0.9011	30.1667 / 0.8727
Proposed	36.9555 / 0.9356	34.6198 / 0.9033	30.4060 / 0.8770

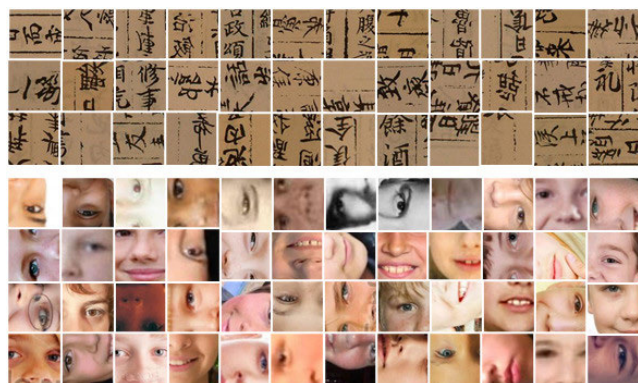


FIGURE 4. A part of HR patch images that have been cropped, rotated, and flipped for training.

ducted experiments to generate text and face SR images using four different methods (vanilla EDSR and three semantic SR methods) and compare them.

To implement our SR method, we used the EDSR model with 16 residual blocks as the SR backbone network in Fig. 1, used the discriminator network of SRGAN [13] as discriminator, and modified the open source code [32] implemented using the TensorFlow library. For comparison, we also implemented the Xu et al.’s method (multi-class GAN) and the Park’s method on our own. In the Xu et al.’s method, we did not have a single GAN generate both text and face SR images, but built two separate GANs for text and face

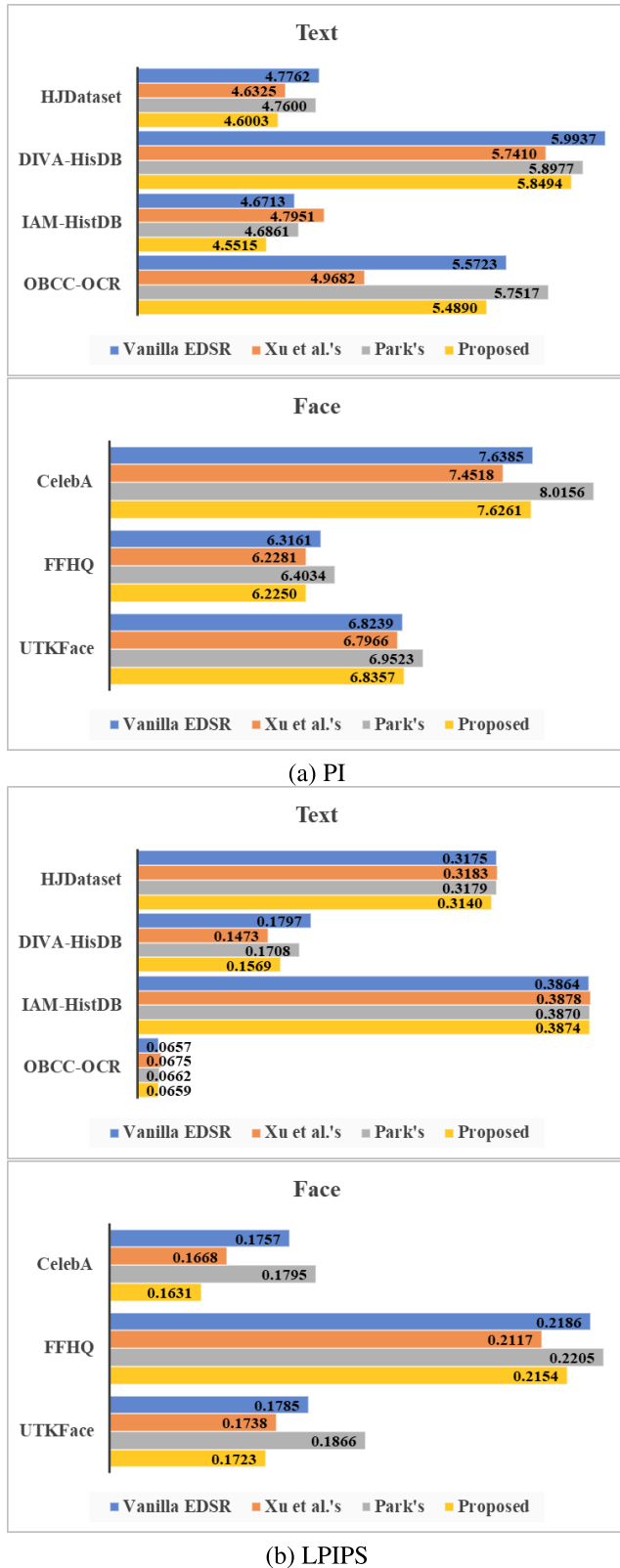


FIGURE 5. Qualitative image quality evaluation of SR images produced by different SR methods.

images; thus, each GAN uses the EDSR model as generator and a single discriminator for identifying real text or face

images. All the networks were trained from scratch using the Adam optimizer with momentum terms $\beta = (0.9, 0.999)$, learning rate = 10^{-4} , upscale factor = 4, batch size = 16, and total steps = 100,000 on a single RTX 3090 GPU. In the Xu et al.'s method, the weighting factors $\lambda_1, \lambda_2, \lambda_3$, and margin α were set to 1, 10^{-3} , 0.1, and 1, respectively. In the Park's method, the semantic loss weight λ was set to 0.006.

B. DATASETS AND EVALUATION METRICS

For text image SR, we used Old book Chinese character OCR (OBCC-OCR) [33] dataset, consisting of 1,108 images, for training and IAM-HistDB [34], DIVA-HisDB [35], and HJDataset [36] datasets, consisting of 127, 120, and 341 images, respectively, for testing. The training and testing datasets contain scanned text images of handwritten or printed historical manuscripts, written in different languages including Chinese, German, English, Latin, and Japanese (see Fig. 2). 10% of the images in the OBCC-OCR datasets were also used for testing.

For face image SR, we used the aligned and cropped images in UTKFace [37] dataset, consisting of 21,398 images, for training and FFHQ [38] and CelebA [39] datasets, consisting of 900 and 200 images, respectively, for testing. The datasets contain face images of people of different ages, genders, and ethnicities (see Fig. 3). 10% of the images in the UTKFace datasets were also used for testing.

The original images of each dataset were used as HR images and they were downsampled using the cubic interpolation with anti-aliasing on to create LR images. The HR images were randomly cropped into patches of size 96×96 , rotated, and flipped for training (see Fig. 4).

To evaluate the quantitative and qualitative quality of SR images, we computed PSNR, structural similarity index measure (SSIM), perception index (PI) [40], and learned perceptual image patch similarity (LPIPS) [41] values. The better the image quality, the higher the PSNR and SSIM values are, but the lower the PI and LPIPS values are.

C. RESULTS AND DISCUSSION

Table 1 shows the PSNR and SSIM values of text SR images produced by different SR methods. The semantic SR methods use the EDSR model as their backbone network, but had higher PSNR and SSIM values than the vanilla EDSR. It means that the semantic priors obtained by each method contributed to improving the quantitative quality of SR images. However, as mentioned in the previous study [5], Xu et al.'s method tended to be over-fitted to the training dataset. As a result, its PSNR and SSIM values were highest for the OBCC-OCR dataset, but lower than Park's method or the proposed method for the other datasets. Xu et al.'s method is similar to the proposed method in that it is a GAN-based framework, but neglects to reduce semantic differences between images belonging to the same category, resulting in reduced generalization ability. The proposed method



FIGURE 6. Visual comparison of text SR images produced by different SR methods. The images were cropped and enlarged to improve visibility.

outperformed the other methods without being over-fitted to the training dataset. The proposed method consistently showed higher PSNR and SSIM values than Park's method for all datasets, indicating that our GAN-based framework is more effective for semantic SR than the VGG-based framework.

Table 2 shows the PSNR and SSIM values of face SR images produced by different SR methods. The noticeable thing was that the Park's method did not work correctly and had lower PSNR and SSIM values than the vanilla EDSR. We think that this is because the face images are semantically less similar compared with the text images and the pretrained VGG network may not be a good solution for extracting semantic information from images. Nevertheless, direct matching (the semantic loss in Park's method) of incomplete semantic information extracted from SR images may cause the SR process to proceed in the wrong way. In contrast, the proposed method showed highest PSNR and SSIM values for all the datasets, in spite of being derived from the Park's method. This is because the proposed method has the discriminator well trained to identify semantically similar images and transfers the semantic prior indirectly to the generator via adversarial learning. This indicates again that our GAN-based framework is effective for semantic SR.

Figure 5 shows the PI and LPIPS values of text and face SR images produced by different SR methods. First, the Park's

method was not helpful for decreasing PI and LPIPS values of text SR images. This means that the Park's method improves only the quantitative quality of SR images (this was not shown in the previous study [5]). As aforementioned, the Park's method did not work correctly in enhancing face SR images; thus, its PI and LPIPS values were consistently higher than the vanilla EDSR. In general, GAN-based methods (i.e., Xu et al.'s and proposed ones) showed good performance in decreasing PI and LPIPS values of SR images. Because their results varied depending on datasets and image categories, it was difficult to judge which was better between the two. However, there was a discernible difference between the perceptual quality of SR images produced by the two methods (see Figs. 6 and 7).

In most results, the visual difference of SR images produced by different SR methods was not clearly observed with the naked eye. So, to clarify the difference, we cropped and enlarged the SR images as shown in Figs. 6 and 7. In the text SR images of vanilla EDSR, the letters were seriously distorted, reducing text readability. However, semantic SR methods reduced the distortion, and in particular, GAN-based methods (Xu et al.'s and proposed) produced visually much better SR images. Among them, the proposed method best reconstructed realistic texture and local details of HR images; thus, its SR images seemed most similar to the corresponding HR images. In the face SR images, the visual difference was mainly observed in the mouth and eyes. The vanilla EDSR

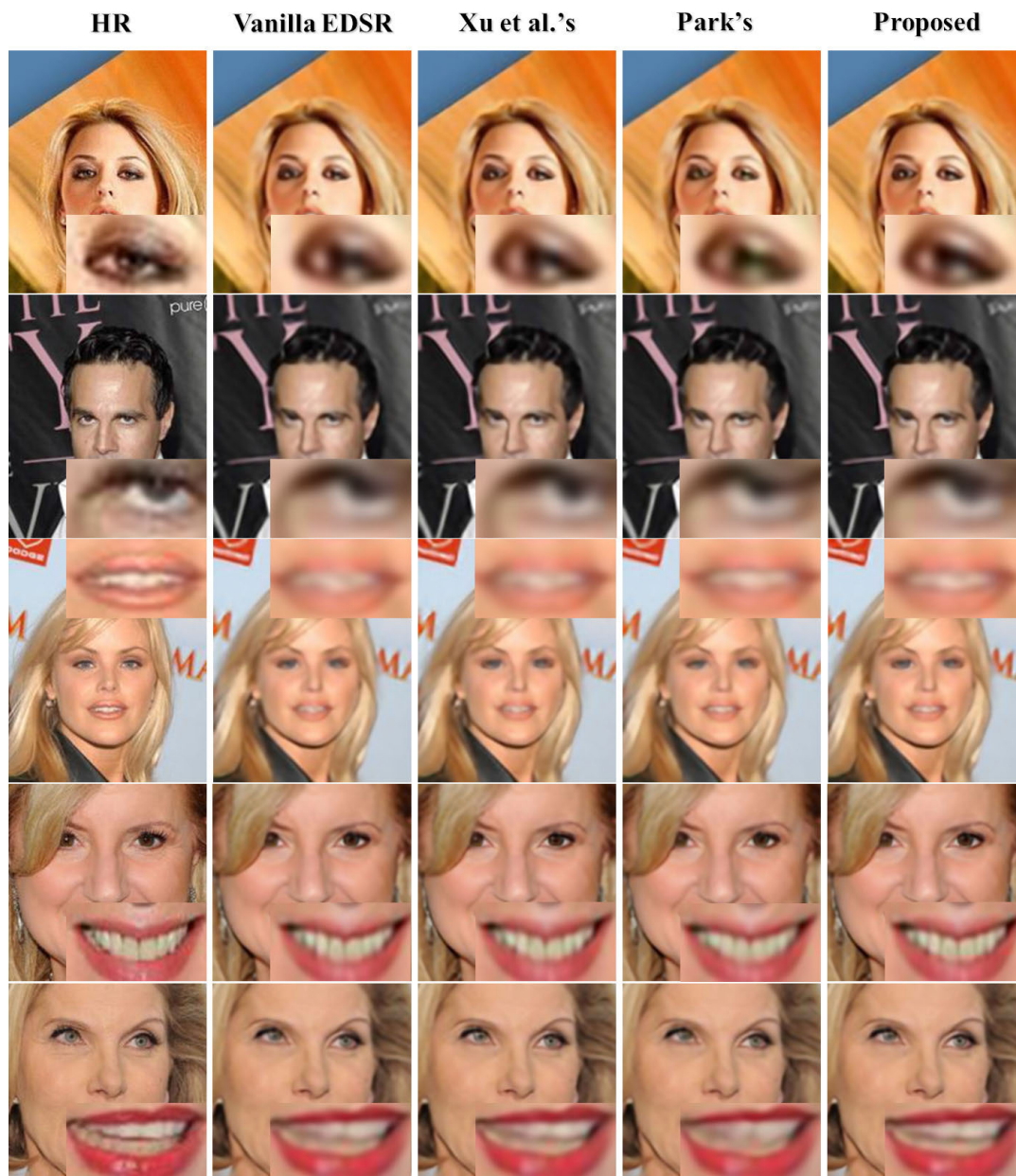


FIGURE 7. Visual comparison of face SR images produced by different SR methods. The images were cropped and enlarged to improve visibility.

and the Park's method sometimes produced unnatural and blurry artifacts (e.g., see the third and last row of Fig. 7). Xu et al.'s method showed good performance in producing visually plausible SR images, but failed to clearly reconstruct local textures. From the results, it can be seen that the local textures and shape of the mouth and eyes were most clearly reconstructed by the proposed method.

From the above results, we can conclude that useful semantic information was extracted by the proposed method and contributed to improving both the quantitative and qualitative qualities of SR images. The improvement was greater than when the other semantic methods were used.

D. LIMITATIONS

The proposed method has yet to fully recover local details, as shown in Figs 6 and 7. This is because the binary discriminator used in the proposed method distinguishes semantic differences between images at the image level; thus, the adversarial loss inevitably has limitations in recovering local details in the pixel level. To address this problem, we will need to design a more sophisticated discriminator. As a preliminary experiment, we tried to use an U-net structure discriminator [42]. It seemed to help reconstruct small textures, but its overall performance was poor compared to the binary discriminator.

It also showed a tendency to be overfitted to the training dataset.

The training parameters in our experiments were not fully optimized. For example, the weight factor (κ) in Eq. 1 were heuristically set. A small κ may cause the semantic information to be less learned, while a large κ may not recover local detailed textures and degrade the SR image quality [5]. Therefore, the parameters need to be fine-tuned for further improvement of performance.

V. CONCLUSION AND FUTURE WORK

In this study, we proposed a semantic SR method that is based on the GAN framework and self-distillation to enable a baseline CNN-based SR model to learn semantic information of images, thereby improving their generalization abilities. The method trained a discriminator to distinguish semantic differences between images belonging to the same category and led the baseline SR model to learn an implicit category-specific semantic prior by minimizing the adversarial loss via self-distillation. Therefore, the proposed method was able to effectively incorporate a decent semantic prior to the SR process without a pretrained network. In experiments with various text and face datasets, the proposed method was able to generate text and face SR images with improved quantitative and qualitative image quality (up to 0.87 dB in PSNR and up to 0.17 in PI) and outperformed existing semantic SR methods.

However, as mentioned in Sec. IV-D, the semantic prior could not be obtained at the pixel level, and the proposed method had a limitation in reconstructing local details. Currently, we are trying to find ways to resolve the problem.

From the perspective of contrastive learning [43], semantically different images (i.e., negative samples) can also help to obtain more robust semantic priors. The related experiment would be an interesting future work.

ACKNOWLEDGMENT

This work utilized “Old book Chinese character OCR” dataset built with the support of the National Information Society Agency and with the fund of the Ministry of Science and ICT of South Korea. The dataset can be downloaded from aihub.or.kr.

REFERENCES

- [1] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, “Image super-resolution: The techniques, applications, and future,” *Signal Process.*, vol. 128, pp. 389–408, Nov. 2016.
- [2] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi, and H. Ghayvat, “CNN variants for computer vision: History, architecture, application, challenges and future scope,” *Electronics*, vol. 10, no. 20, p. 2470, Oct. 2021.
- [3] Z. Wang, J. Chen, and S. C. H. Hoi, “Deep learning for image super-resolution: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2021.
- [4] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadel, M. Al-Amidie, and L. Farhan, “Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions,” *J. Big Data*, vol. 8, no. 1, Mar. 2021, Art. no. 53.
- [5] H. Park, “Semantic super-resolution of text images via self-distillation,” *Electronics*, vol. 11, no. 14, p. 2137, Jul. 2022.
- [6] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.
- [7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. ICLR*, 2015, pp. 1–14.
- [8] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.
- [9] T.-B. Xu and C.-L. Liu, “Data-distortion guided self-distillation for deep neural networks,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 5565–5572.
- [10] S. Yun, J. Park, K. Lee, and J. Shin, “Regularizing class-wise predictions via self-knowledge distillation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13873–13882.
- [11] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in *Proc. ECCV*, 2014, pp. 184–199.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [13] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.
- [14] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. R. Fu, “Image super-resolution using very deep residual channel attention networks,” in *Proc. ECCV*, Sep. 2018, pp. 294–310.
- [15] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Proc. ECCV*, vol. 9906, 2016, pp. 694–711.
- [16] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, “ESRGAN: Enhanced super-resolution generative adversarial networks,” in *Proc. ECCV Workshops*, Sep. 2019, pp. 63–79.
- [17] Y. Choi and H. Park, “Improving ESRGAN with an additional image quality loss,” *Multimedia Tools Appl.*, vol. 82, no. 2, pp. 3123–3137, Jan. 2023.
- [18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16 × 16 words: Transformers for image recognition at scale,” in *Proc. ICLR*, 2021, pp. 1–21.
- [19] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using Swin transformer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.
- [20] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, “Activating more pixels in image super-resolution transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22367–22377.
- [21] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, “Transformer for single image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 456–465.
- [22] Z. Zhou, G. Li, and G. Wang, “A hybrid of transformer and CNN for efficient single image super-resolution via multi-level distillation,” *Displays*, vol. 76, Jan. 2023, Art. no. 102352.
- [23] G. Gendy, G. He, and N. Sabor, “Lightweight image super-resolution based on deep learning: State-of-the-art and future directions,” *Inf. Fusion*, vol. 94, pp. 284–310, Jun. 2023.
- [24] J. Sun, J. Zhu, and M. F. Tappen, “Context-constrained hallucination for image super-resolution,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 231–238.
- [25] R. Timofte, V. De Smet, and L. Van Gool, “Semantic super-resolution: When and where is it useful?” *Comput. Vis. Image Understand.*, vol. 142, pp. 1–12, Jan. 2016.
- [26] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, “Learning to super-resolve blurry face and text images,” in *Proc. ICCV*, Oct. 2017, pp. 251–260.
- [27] X. Wang, K. Yu, C. Dong, and C. Change Loy, “Recovering realistic texture in image super-resolution by deep spatial feature transform,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 606–615.
- [28] T. Frizza, D. G. Dansereau, N. M. Sersht, and M. Bewley, “Semantically accurate super-resolution generative adversarial networks,” *Comput. Vis. Image Understand.*, vol. 221, Aug. 2022, Art. no. 103464.

- [29] C. Chen, X. Shi, Y. Qin, X. Li, X. Han, T. Yang, and S. Guo, "Real-world blind super-resolution via feature matching with implicit high-resolution priors," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 1329–1338.
- [30] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, "Learning the degradation distribution for blind image super-resolution," 2022, *arXiv:2203.04962*.
- [31] Y. Zhang, H. Wang, C. Qin, and Y. Fu, "Learning efficient image super-resolution networks via structure-regularized pruning," in *Proc. ICLR*, 2022, pp. 1–12.
- [32] *Single Image Super-Resolution With EDSR, WDSR and SRGAN*. Accessed: Oct. 4, 2023. [Online]. Available: <https://github.com/krasserm/super-resolution>
- [33] *Old Book Chinese Character OCR Data*. Accessed: Oct. 4, 2023. [Online]. Available: <https://aihub.or.kr>
- [34] *IAM Historical Document Database*. Accessed: Oct. 4, 2023. [Online]. Available:
- [35] *DIVAHISDB Dataset of Medieval Manuscripts*. Accessed: Oct. 4, 2023. [Online]. Available: <https://www.unifr.ch/inf/diva/en/research/software-data/diva-hisdb.html>
- [36] *HJDataset: A Large Dataset of Historical Japanese Documents with Complex Layouts*. Accessed: Oct. 4, 2023. [Online]. Available: <https://dell-research-harvard.github.io/HJDataset/>
- [37] *UTKFace Dataset*. Accessed: Oct. 4, 2023. [Online]. Available: <https://susnqq.github.io/UTKFace/>
- [38] *Flickr-Faces-HQ Dataset*. Accessed: Oct. 4, 2023. [Online]. Available: <https://github.com/NVLabs/ffhq-dataset>
- [39] *CelebFaces Attributes Dataset*. Accessed: Oct. 4, 2023. [Online]. Available: <https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>
- [40] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," in *Proc. ECCV Workshops*, Sep. 2018, pp. 334–355.
- [41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
- [42] E. Schönfeld, B. Schiele, and A. Khoreva, "A U-Net based discriminator for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8204–8213.
- [43] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.



HANHOON PARK received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from Hanyang University, Seoul, South Korea, in 2000, 2002, and 2007, respectively. From 2008 to 2011, he was a Postdoctoral Researcher with NHK Science and Technology Research Laboratories, Tokyo, Japan. In 2012, he joined the Department of Electronic Engineering, Pukyong National University, Busan, South Korea, where he is currently a Professor. His current research interests include augmented reality, human–computer interaction, and deep learning application.

• • •