IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 27, 2019, accepted May 20, 2019, date of publication June 10, 2019, date of current version June 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921859

Component Semantic Prior Guided Generative Adversarial Network for Face Super-Resolution

LU LIU, SHENGHUI WANG, AND LILI WAN
Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China

Corresponding authors: Shenghui Wang (shwang@bjtu.edu.cn) and Lili Wan (llwan@bjtu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61572064.

ABSTRACT Face super-resolved (SR) images aid human perception. The state-of-the-art face SR methods leverage the spatial location of facial components as prior knowledge. However, it remains a great challenge to generate natural textures. In this paper, we propose a component semantic prior guided generative adversarial network (CSPGAN) to synthesize faces. Specifically, semantic probability maps of facial components are exploited to modulate features in the CSPGAN through affine transformation. To compensate for the overly smooth performance of the generative network, a gradient loss is proposed to recover the high-frequency details. Meanwhile, the discriminative network is designed to perform multiple tasks which predict semantic category and distinguish authenticity simultaneously. The extensive experimental results demonstrate the superiority of the CSPGAN in reconstructing photorealistic textures.

INDEX TERMS Facial component, face super-resolution, generative adversarial networks, multiple task, semantic prior.

I. INTRODUCTION

Face super-resolution (SR) task refers to reconstructing a high-resolution (HR) image from a low-resolution (LR) facial image. It increases high-frequency details and removes degradation due to various factors including blur, noise and low-resolution caused by the imaging acquisition device. Generating photo-realistic super-resolved faces is beneficial for a series of face-related tasks, including face attribute recognition [1], face alignment [2], and face recognition [3] in complex real-world scenarios. It has attracted a large amount of attention in image processing and computer vision communities [4]–[6].

Face SR is a domain specific reconstruction problem. While there are numerous algorithms for performing face SR, the majority of them are devoted to generic images [4], [7], [8]. These methods cannot fully utilize the characteristics of facial images with their structural geometry and similar appearance [9] and thus cause a loss of fine-grained details. To solve this issue, different kinds of face-specific information are adopted as latent priors to guide the super-resolved generating process. Among them, most are spatial location related. Zhu *et al.* utilized a facial correspondence field to describe the spatial configuration and corresponding properties [10]. Yu *et al.* employed facial component heatmaps to explicitly incorporate structural information of face into the synthesizing process [11]. On the basis of different priors, the overall visual quality of reconstruction are significantly improved.

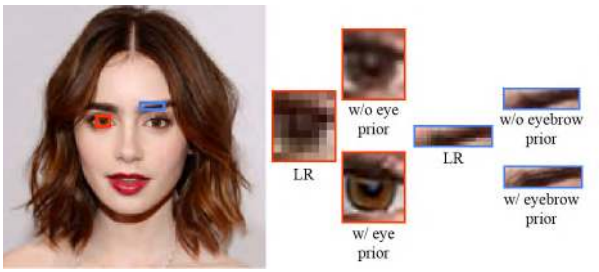

FIGURE 1. Exemplar of synthesized images with or without semantic guidance. Training a CNN-based generative network with a guidance of semantic prior can add convincing details which fall into their underlying classes. Owing to the semantic information, pupil and brow patches can be recovered clearly.

Though great strides have been made, generating photo-realistic textures is still a challenging problem. Given that the results obtained with spatial prior cannot generate textures which are faithful to their natural classes, the semantic prior is utilized and demonstrated to be non-trivial in general SR problem [9], [12]. As shown in Fig. 1, with the guidance of semantic information, super-resolved patches appear

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

perceptually convincing. Driven by this idea, we propose semantic segmentation probability maps to guide the process of reconstruction. These maps are a set of stacked probability values which are able to represent the semantic categories of well-segmented facial components. With the guidance of the strong semantic prior, the proposed component semantic prior (CSP) layer can modulate the intermediate features of network to synthesize photo-realistic facial details.

Meanwhile, considering that the face SR problem is a ill-posed one, an effective way to constrain the solution space of convolution neural networks (CNN) based methods is through learning mapping loss functions from LR and HR exemplars [13]–[18]. Pixel-wise errors, such as the mean squared error (MSE) and the mean absolute error (MAE), are the most widely applied losses. Though they are helpful in improving the peak signal-to-noise (PSNR) score, they have shortcomings to capture perceptually relevant differences and thus encourage a generation of blurry and overly-smooth results.

To remedy the deficiency, a new gradient loss is introduced to recover high-frequency details. In order to reconstruct them with low-frequency information at the same time, a mask is utilized to separate them on the basis of the gradient magnitude of an image. We seek to constraint the generative network with the gradient loss in the region of high-frequency textures. Meanwhile, a perceptual loss, which relies on some pre-trained models [19], is utilized to impact the feature space instead of pixel space of low-frequency regions. The weighted combination of two losses can capture more high-frequency details, *e.g.* the hair, while maintaining the perceptual fidelity of the original HR face images.

Moreover, inspired by recent success of Generative Adversarial Networks (GANs) based methods in synthesizing images, Yu *et al.* first developed a GANs-based face SR algorithm to reconstruct faces. The idea behind GANs is to train a generative network G to fool a discriminative network D, which is interactively trained to distinguish super-resolved faces from real ones. The network D is utilized to optimize the smoothness of images which are synthesized by the only generative network. In this paper, we not only inherit plain D to distinguish the validity of facial images, but also to predict the semantic categories of the input. Multiple tasks of D make the process of reconstruction more robust in poor conditions.

In summary, the main contributions of this paper include:

- We propose a novel CSPGAN to generate photo-realistic details in face super-resolution. To the best of our knowledge, this is the first component semantic prior guided face SR method.
- In CSPGAN, we design a new gradient loss to capture high-frequency information and concatenate it with the perceptual loss to generate satisfying facial textures.
- In CSPGAN, we design a multiple task discriminative network to distinguish authenticity and predict semantic category simultaneously.

II. RELATED WORK

Exploiting facial related priors is the key factor that distinguishes face SR problem from generic SR tasks. In this section, we will review some prior related works on face super-resolution methods. As for different kinds of priors, the existing face SR methods can be classified into three categories: (i) global-based methods [20]–[24], (ii) patch-based methods [25]–[31], and (iii) learning-based methods [32]–[42].

The global-based methods employ holistic priors to represent the input LR faces and synthesize SR ones as weighted combinations of training samples. Wang *et.al* employed principal component analysis (PCA) algorithm to model the global prior variations of facial appearance in the eigenspace [20]. Baker and Kanade proposed pyramid-based algorithm to learn a prior on the distribution of face gradients [43]. However, the performance of global-based methods depends on whether the training set is distributed widely and may degrade greatly if the number of training samples is insufficient.

Local feature priors are utilized to compensate the details of generated SR faces in patch-based methods. In these methods, facial image is considered as a highly structured object and is divided into patches according to the position prior. Assuming the image patches share the same local geometry, the position information can be served as a constraint to map the LR patches to HR ones through combinations. Liu *et al.* proposed a two-step approach which decomposed the SR process into global reconstruction and local residual compensation [44]. Motivated by this work, plenty of patch-based algorithms were proposed to solve the problem through implicitly coding by representing the input LR patches locally [25], [28], [45], collaboratively [26], and sparsely [29], [38] or explicitly regression [46]–[48]. Nonetheless, a local patch prior is insufficient to infer the holistic structure of facial images, especially when the upscaling factor is too large.

In the meantime, as a domain specific SR technique, the face SR method is influenced by the rapid development of deep learning techniques. With the help of a large training dataset, learning-based methods aim to predict super-resolved faces through LR and HR facial image pairs [49]. The prior knowledge is learned from a set of HR training images and then used to reconstruct SR images. Due to larger magnification factor that learning-based methods can achieve, they attract more attention in recent research for practical applications. Yu *et al.* [17] localized facial components in the LR images by practicing a facial landmark detector and then reconstructed missing high-frequency details from similar HR references. In their remarkable work, the facial components need to align accurately and the performance degrades dramatically when misalignment occurs. In this paper, instead of enforcing a precise alignment on input facial images, we preserve the spatial information and assign

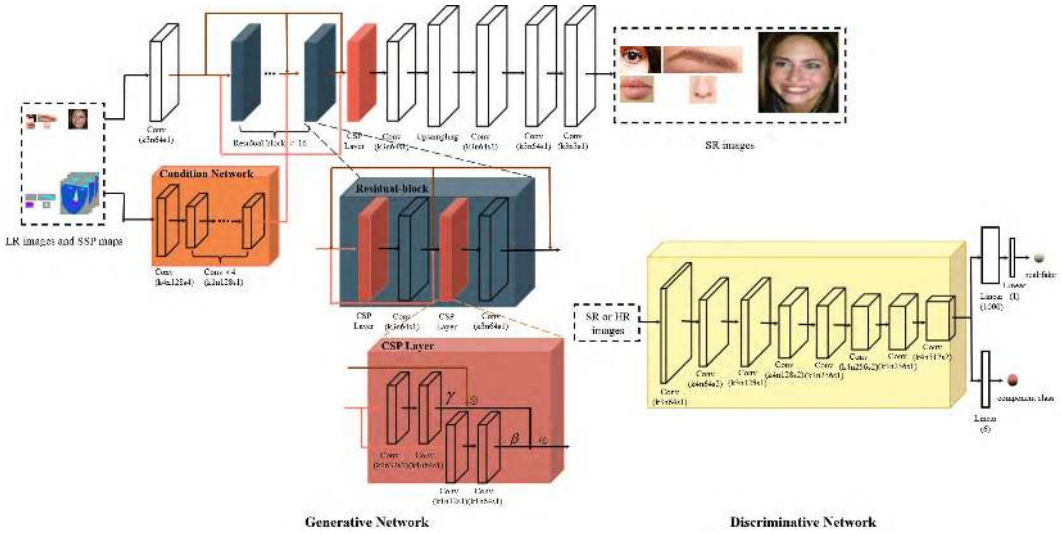

FIGURE 2. The architecture of CSPGAN network. LR, SR, and HR images represent the low-resolution, super-resolution and high-resolution facial images.

semantic categories' probabilities as prior. They are not only more robust to miner misalignment, but also hold richer information to assist reconstruction.

III. METHOD

In face SR problem, given a low-resolution facial image I_{lr} as input, the super-resolution image I_{sr} is generated by the generator G where I_{hr} is the corresponding ground truth. High-resolution images are only available during training while low-resolution ones are obtained by a downsampling operation with factor r. The reconstruction process can be represented as G parameterized by θ:

$$I_{sr} = G(I_{lr}|\theta). \tag{1}$$

Our goal is to find a suitable set of parameters for the generator G during training. For a given LR input I_{lr} and its corresponding HR counterpart I_{hr}, we solve:

$$\hat{\theta} = \arg\min_{\theta} \sum \mathcal{L}(G(I_{lr}), I_{hr}). \tag{2}$$

In the following, we will introduce our CSPGAN in details.

A. NETWORK ARCHITECTURE

As shown in Fig. 2, our CSPGAN is composed of two parts: a generative network G and a discriminate network D. In network G, the semantic segmentation probability (SSP) maps are created to preserve the semantic information of facial components. Instead of reconstructing facial components independently [50], SSP maps are utilized as feed-forward guidances to alter the behavior of G through a well-designed component semantic prior (CSP) layer. Trained with the end-to-end strategy, CSP layers affinely amend intermediate features of G on the basis of SSP maps. Specifically, to further share the parameters, we use a small condition network to generate shared values for broadcasting thoroughly. Meanwhile, the discriminator D has the ability of multi-tasking to not only determine authenticity, but also predict the semantic category for the inputs. Different from the previous discriminator, our D encourages generator G to reconstruct more confident results.

B. DETAILS INSIDE GENERATIVE NETWORK G

In CSPGAN, the generator G is a feed-forward network that modulates features according to the semantic information of facial components. It consists of four parts: semantic segmentation probability maps, component semantic prior layer, condition network, and residual block.

1) SEMANTIC SEGMENTATION PROBABILITY MAPS

As mentioned before, the issue with existing face SR methods is easily trapped to degraded details which are visually identical between different semantic sections, *e.g.* the eye and eyebrow. One of the solutions is to fulfill the synthesizing process with effective semantic information. An effective way is represented as semantic category along with its spatial location. In our generator G, a semantic segmentation probability maps P is introduced to represent them. Specifically,

$$\Phi = P = (P_1, \cdots, P_k, \cdots, P_K), \tag{3}$$

where P_k represents the probability map of K^{th} category and K is the total number of considered categories.

For the ground truth I_{hr}, we use Openface [51] to directly perform face detection and obtain the landmark localizations. Then we initialize K stacked maps of probabilities to zero which have the same size of inputs. Based on the k^{th} semantic category that each pixel belongs to, the value of same location in k^{th} SSP map is set to a random value in [0.8, 1] while values of other $k-1$ maps are set to values in [0, 0.2]. We intentionally avoid the hard probability values of ones

and zeros for robustness. Note that the SSP maps are easy to truncate, random flip, or rotate along with HR faces.

With the expression of both location and sematic prior information in SSP maps, the process of face SR in Eq. (1) can be expressed as:

$$I_{sr} = G(I_{lr}, \Phi|\theta). \tag{4}$$

2) COMPONENT SEMANTIC PRIOR LAYER

The motivation of the component semantic prior layer is to change the behavior of the generative network on the basis of semantic priors to synthesize super-resolved faces. In this section, we show a direct way to alternate the behavior of G. This feed-forward technique adaptively influences the outputs rapidly.

To be specific, we design a CSP layer to learn a mapping function C which outputs a set of parameter pairs (γ, β) based on semantic prior Φ. Through the learned parameters, CSP layers are able to influence the synthesized SR faces by applying an affine transformation at intermediate feature values of G. As seen in Fig. 2, the CSP layers are directly embedded into G. During testing, only a single forward pass is needed to generate the SR facial images, given the LR inputs and SSP maps. The process can be described as follows:

A pair of affine transformation parameters (γ, β) is modeling the prior Φ through a mapping function C,

$$(\gamma, \beta) = C(\Phi). \tag{5}$$

Consequently, the target SR faces in Eq. 4 can be calculated by:

$$I_{sr} = G(I_{lr}, (\gamma, \beta)|\theta). \tag{6}$$

By scaling and shifting intermediate features, the CSP layer performs affine transformation in generator G after obtaining γ and β. The process is as follows:

$$CSP(F|\gamma, \beta) = \gamma \otimes F + \beta, \tag{7}$$

where F denotes the features of input facial images in G with the same dimension as γ and β, and $\otimes$ refers to the Hadamard product of element-wise multiplication. Most of the transformations are calculated in the LR space and followed with the upsampling operation that broadcasts the computation thoroughly. The details of CSP layer is shown in Fig. 2.

3) CONDITION NETWORK

To share the semantic prior that SSP maps contain, a conditional network plays a role of delivering conditions to all the CSP layers. This small network is filled with convolutional operations. Meanwhile, we still keep few parameters inside each CSP layer to further adapt the shared conditions to the specific parameters γ and β, providing fine-grained control of the features.

4) RESIDUAL BLOCK

Different from other face SR methods, we introduce residual blocks embedded with two CSP layers to change the generative network's behavior. Skip connection [52] is used to ease the training of deep CNN-based generative networks. On top of them, several upsampling layers are placed to magnify the input LR facial images into their high-resolved sizes.

C. DETAILS INSIDE GENERATIVE NETWORK D

With a delicate design, the discriminator D undertakes two different tasks: determine whether the synthesized image is real or fake, and estimate the sematic categories which the components belong to (see Fig. 2). We apply a CNN-based network and use LeakyReLU activation ($\alpha = 0.2$) and avoid max-pooling throughout the network. Strided convolutions are employed in the intermediate layer for D to gradually decrease the dimensions. Although the inputs are different sizes, they use similar convolution layers rather than fully connected layers since the training images are cropped to contain only one category. This restriction is not applied on test images. We find this strategy facilitates the generation of images with realistic textures and robust characteristics.

IV. LOSS FUNCTION

The loss function plays an important role in our generative network G. While $\mathcal{L}_G$ is commonly modeled by the MSE [53], [54], we design a loss function that helps assess perceptually relevant characteristics and preserve high-frequency details. We formulate the loss as the weighted sum of a perceptual loss, a gradient loss and an adversarial loss:

$$\mathcal{L}_G = \underbrace{\lambda_{per}\mathcal{L}_{per} + \lambda_{grad}\mathcal{L}_{grad}}_{content\ loss} + \underbrace{\lambda_{adv}\mathcal{L}_{adv\_G}}_{adversarial\ loss}. \tag{8}$$

In the following, we describe the perceptual loss and the gradient loss as two components of content loss. The details of each loss are defined below.

A. CONTENT LOSS

Optimizing a pixel-wise loss may cause perceptually blur and a lack of high-frequency information. In this section, we design a content loss which is a weighted combination of a perceptual and a gradient loss to address this issue. By training end-to-end, the content loss is performing perfectly in reconstruction.

1) PERCEPTUAL LOSS

Following Ledig's work [52], a perceptual loss is utilized to help assess perceptually relevant characteristics by using high-level features of a pre-trained VGG-19 network [55]. In details,

$$\mathcal{L}_{per} = \|\phi(I_{sr}) - \phi(I_{hr})\|^2, \tag{9}$$

where ϕ denotes the pre-trained VGG model. Similar to [4], we use the feature maps obtained by the 'relu5_3' layer and compute the MSE on the feature activations.

2) GRADIENT LOSS

Drawing an inspiration from the common fact that high-frequency details lie under high gradient pixels, we propose

a gradient loss to recover the image gradients while rescuing the missing high-frequency textures. Intuitively, the gradient loss is formulated as [59]:

$$\mathcal{L}_{grad} = \|\nabla_x I_{hr} - \nabla_x I_{sr}\|_1 + \left\|\nabla_y I_{hr} - \nabla_y I_{sr}\right\|_1, \quad (10)$$

where $\nabla_x I_{hr}$ and $\nabla_y I_{hr}$ denote the directional gradients of I_{hr} along the horizontal (denoted by x) and vertical (denoted by y) directions, respectively.

It is noted that minimizing the gradient loss will help to recover the gradients, but it will cost a degradation of reconstruction performance. By introducing a mask M to separate the high-frequency part from the low-frequency one, the constraint on gradient-level will not affect the pixel-level. Therefore, optimizing the weighted joint loss is able to preserve both low-frequency content and high-frequency structure of facial images. Specifically, the mask M is used to decompose the image I by

$$I = M \odot I + (1 - M) \odot I, \quad (11)$$

where $M_{i,j} \in [0, 1]$. Given the gradient magnitude GM, where $GM_{i,j} = \sqrt{(\nabla_x I_{i,j})^2 + (\nabla_y I_{i,j})^2}$, we can define the mask M as the normalization of GM into $[0, 1]$:

$$M = (GM - \min(GM))/(\max(GM) - \min(GM)), \quad (12)$$

where $\max(GM)$ and $\min(GM)$ denote the maximum and minimum value in GM, respectively. Finally, we define our content loss as:

$$\begin{aligned}\mathcal{L}_{content}(I_{sr}, I_{hr}) = \lambda_{grad}\mathcal{L}_{grad}(M \odot I_{sr}, M \odot I_{hr}) \\ + \lambda_{per}\mathcal{L}_{per}((1 - M) \odot I_{sr}, (1 - M) \odot I_{hr}), \quad (13)\end{aligned}$$

where $\odot$ denotes the element-wise multiplication.

B. ADVERISARIAL LOSS

Our framework is based on adversarial learning. It consists of a generator G and a discriminator D, which are parameterized by θ and η respectively. They are jointly trained on the basis of the objective function, which can be written as:

$$\min_{\theta} \max_{\eta} \mathbb{E}_{y \sim P_{I_{hr}}} \log D(y|\eta) + \mathbb{E}_{x \sim P_{I_{lr}}} \log(1 - D(G(x|\theta)|\eta)), \quad (14)$$

where $x \sim P_{I_{lr}}$ and $y \sim P_{I_{hr}}$ are LR and HR samples' empirical distributions.

For generator G, the adversarial loss can be written as:

$$\mathcal{L}_{adv\_G} = \mathbb{E}_{x \sim P_{I_{lr}}}[(D(G(x\,|\theta)\,|\eta) - 1)^2]. \quad (15)$$

Meanwhile, the training process alternately minimizes the objective function of discriminator D as follows:

$$\mathcal{L}_D = \mathbb{E}_{x \sim P_{I_{lr}}}[(D(G(x\,|\theta)\,|\eta))^2] + \mathbb{E}_{x \sim P_{I_{hr}}}[(D(x\,|\eta) - 1)^2]. \quad (16)$$

V. EXPERIMENTS

A. DATASET

In this paper, experiments are evaluated on Labeled Faces in the Wild (LFW) for its diversity in facial images, such as expression, occlusion, aging, etc. We assume five primary categories, *i.e.*, eyebrow, eye, nose, lip, and facial. A 'background' category is used to express regions that do not appear in the aforementioned categories. For LFW dataset, we use 10023 images for training, and 1091 images for evaluation.

In addition, we collect a new set of high-resolution images to complement the facial details since the resolution of LFW images is still quite low. By querying the Google search engines using 'high resolution image' and the defined categories as keywords, we gather 573 HR facial images and 349 facial components images. It is called Facial HR images Online (FHRO) dataset.

Following [52], all experiments are performed with scaling factor of $\times 4$ and $\times 8$ between LR and HR facial images. During training, we use MATLAB bicubic and near kernel to downsample HR faces and obtain LR faces. The sizes of cropped HR and LR sub-images are 96×96 and 24×24, respectively.

B. SETTINGS

We initialize the network by parameters pre-trained with perceptual loss and GAN loss on LFW dataset. After initialization, we fine-tune our CSPGAN network on the FHRO dataset on the basis of SSP maps. During training, the size of each mini-batch is set to 16. For optimization, we use Adam [60] with $\beta_1 = 0.9$. The learning rate is set to $1e^{-4}$ and decays by a factor of 2 every 100k iterations. The trade-off parameters λ_{adv}, λ_{per}, and λ_{grad} are empirically set to $5e^{-3}$, 1, and $1e^{-3}$. Alternatively optimizing the generator G and discriminator D, the model usually converges at about $5e^3$ iterations.

In order to evaluate the performance of the network, we compare the synthesized SR facial images to the state-of-the-art methods with qualitative and quantitative analysis. The Peak Signal-to-Noise Ratio (PSNR), structural similarity (SSIM), and feature similarity (FSIM) [61] are employed as our evaluation measurements.

C. COMPARISON WITH STATE-OF-THE-ART METHODS

We compare the CSPGAN with four state-of-the-art SR methods, including FSRCNN [56], ESPCN [57], SRResNet [58], and SRGAN [52]. For FSRCNN and ESPCN, we train the released codes with the same LFW dataset. For SRResNet, we implement 16 residual blocks which take a standard feed-forward convolution network and add skip connections that bypass a few layers. As an improvement of SRResNet, SRGAN adds GANs' structures with perceptual loss function to recover the fine texture details. In the comparison experiments, we train all five methods for qualitative comparisons, but only SRResNet and SRGAN are for quantitative estimation. All the comparative experiments are fine-tuned with the same settings as ours.

FIGURE 3. Comparison between different SR approaches with downsampling factor ×4 in LFW dataset: FSRCNN [56], ESPCN [57], SRGAN [52], SRResNet [58], our proposed CSPGAN and the original HR image.

FIGURE 4. Details of comparison between generated facial images by SRResNet [58], SRGAN [52], CSPGAN w/o $\mathcal{L}_{grad}$, and CSPGAN with magnified factor ×8 in FHRO dataset. The first and third row are restored facial images through different methods while the second and forth row are specific corresponding details of the framed patches. (Zoom in for best view).

First, we compare CSPGAN with the state-of-the-arts methods qualitatively. As shown in Fig. 3, the first two methods cannot recover facial details accurately. They suffer from edge overlaps and blob-like artifacts. SRGAN and SRResNet methods notably improve the high-frequency details, however, they tend to generate monotonous textures. In contrast to the above approaches, our network benefits from the sematic priors of facial components, generating realistic textures, clear outlines, and pleasant colors.

Second, we assess the performances quantitatively by comparing the evaluations on the scale factor ×8 with the test dataset of FHRO. The results are shown in Fig. 4 and Table 1. Although the scores of SRResNet are pretty high, the synthesized facial images are perceptually too smooth and lack of convincing results. Benefitting from the facial semantic prior knowledge, our network with or without gradient loss is all able to accurately reconstruct details that belong to their semantic categories and generate photo-realistic final results. CSPGAN network significantly outperforms state-of-the-arts methods in PSNR, SSIM, and FSIM and recovers more high-frequency details.

D. ABLATION STUDY

To validate the effectiveness of the proposed model for face SR problem, we study our CSPGAN which is converged with only perceptual loss except for the last subsection.

TABLE 1. PSNR, SSIM and FSIM scores of compared methods for ×8 upscaling face super-resolution in FHRO dataset.

Assessment	Bicubic	SRResNet [59]	SRGAN [53]	CSPGAN w/o $\mathcal{L}_{grad}$	CSPGAN
PSNR(dB)	27.887	29.0657	28.9482	29.0205	29.0532
SSIM	0.8182	0.8275	0.8096	0.8105	0.8255
FSIM	0.8605	0.9076	0.8889	0.9218	0.9236

FIGURE 5. Effects of CSPGAN with or without certain facial component's category in the progress of reconstruction. When not labeling a semantic category, *e.g.* the 'eye' component, the generated facial images with magnified factor ×8 will suffer blurry boundaries. The images of the upper row are integral reconstructed faces with their PSNR scores and images of the lower row are the details of the corresponding boxed patches.

Individually, the effects of the proposed gradient loss of the model is demonstrated in the last subsection.

1) EFFECTS OF FACIAL COMPONENTS' CATEGORIES

The facial semantic prior plays an significant role in our generative network. By pruning the facial components' semantic categories in SSP maps, we design an experimental procedure to train CSPGAN separately to observe the influence of different categories. Based on the baseline network, we present a comparison of SSP maps with or without eye and eyebrow categories in Fig. 5. As we can see, the model using full prior information outperform the crippled ones with the PSNR deficiency of 1.1 dB. The outline of synthesized face is blurry in the models without eye or eyebrow.

2) EFFECTS OF CSP LAYERS

A powerful prior representation of facial image may lead to accurate guidance in the network. Here, we mainly focus on how the CSP layers influence the behavior of generator G on the basis of semantic prior. Since the CSP layers are warped in residual block, we test the number of blocks by intuitively tuning $n = 2/4/8/16$ to estimate the effect of CSP layer. We show the details of generated images in Fig. 6. It can be observed that using more CSP layer leads to a deeper structure, a growth of the learning ability of G, and hence better performance. Clearly, synthesized facial images with 16 residual modules have clear boundaries and more convincing textures. Meanwhile, the PSNR and SSIM scores, which are shown in Table 2, verify the upper conclusion.

3) EFFECTS OF MULTIPLE TASKS IN DISCRIMINATOR D

Although it is natural to believe that a plain discriminator is beneficial to judge authenticity and to recover faces, there is still an improved performance of our discriminative network

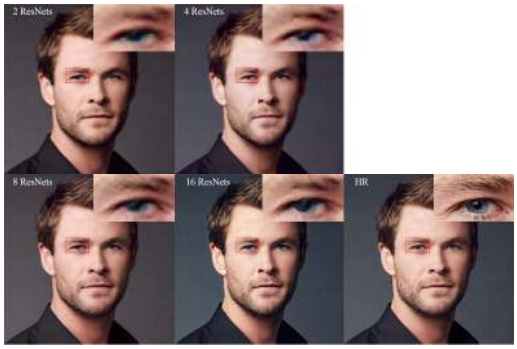

FIGURE 6. Effects of CSPGAN with different number of CSP Layers. The number of CSP layers has a strong influence on the reconstruction ability. Since they are embedded into the residual blocks, we prune the number of ResNets to test it. When $n = 16$, the synthesized face appears most like the ground truth HR image. The upper right legends are the details of framed corresponding patches. (Zoom in for best view).

TABLE 2. PSNR, SSIM and FSIM scores of generator G with different CSP layer number n.

Assessment	$n = 2$	$n = 4$	$n = 8$	$n = 16$
PSNR(dB)	26.7818	27.5874	28.7726	29.3658
SSIM	0.7828	0.8029	0.8167	0.8232
FSIM	0.8820	0.9105	0.9197	0.9244

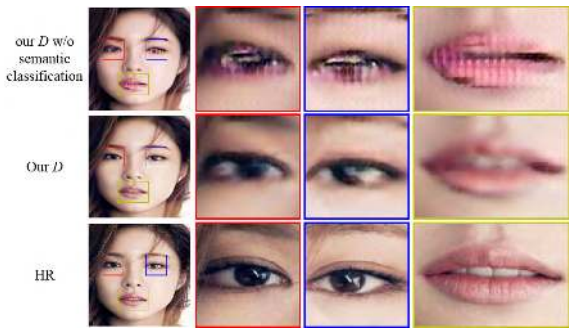

FIGURE 7. Effects of the discriminator D with or without the ability of multi-tasking. Our proposed multi-tasking discriminative network is capable of generating richer and more realistic textures than the plain one. By exploiting same structure of generative network, the first and second row are the illustrations of faces and the corresponding boxed patches. They are generated by discriminative networks which are able to ignore or judge the semantic categories of generated images. The third row shows the ground truth high-resolution images.

to reconstruct photo-realistic textures. As a comparison, our D is able to distinguish which region belongs to the facial component, *e.g.* eye, nose, or lip. In Fig. 7, the extracted patches from generated facial image through generic discriminator D appear to be warped. Using the same structure of generator G, but lack of distinguishing category ability will synthesize details that are not faithful to the underlying class or blob-like artifacts. Our D with multiple tasks can

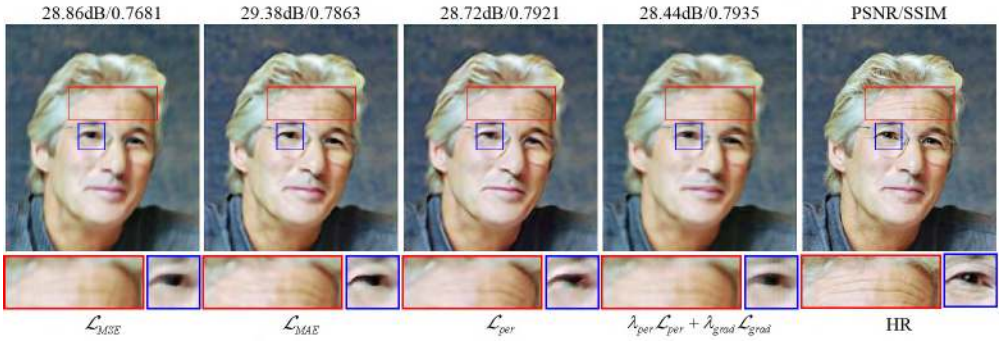

FIGURE 8. Performance comparison of different loss functions. The PSNR and SSIM values are shown above the images. Our proposed loss, a weighted combination of $\lambda_{per}\mathcal{L}_{per} + \lambda_{grad}\mathcal{L}_{grad}$, outperforms other losses in term of perceptual quality.

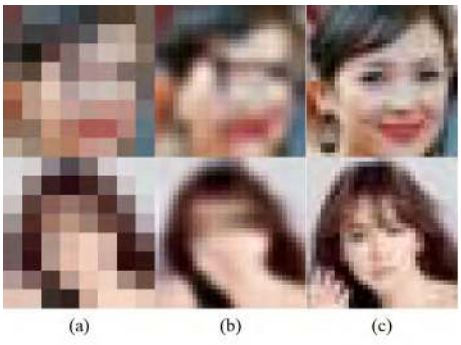

FIGURE 9. Illustration of the limitation with downsampling factor ×4. (a) 8 × 8 low-resolution image. (b) Super-resolved images synthesized by CSPGAN. (c) Original 32 × 32 high-resolution image.

assure G to generate more realistic details and more robust to noises.

4) EFFECTS OF GRADIENT LOSS

To show the effectiveness of the proposed gradient loss, we perform face super-resolution results in $\times 8$ upscaling to study the impacts of the differences. By comparing the mean square error, mean absolute error, perceptual loss, and the proposed gradient loss, CSPGAN is trained under the same settings. As shown in Fig. 8, the proposed content loss of perceptual and gradient loss is able to converge to the highest PSNR score among all the compared experiments. From the reconstructed RGB facial images, we observe generator G trained with $\lambda_{per}\mathcal{L}_{per} + \lambda_{grad}\mathcal{L}_{grad}$ loss is able to capture high-frequency details and maintain the perceptual fidelity of the original HR images.

VI. CONCLUSION, LIMITATION, AND FUTURE WORK

We present a novel face super-resolution method, named CSPGAN, which generates abundant texture and clear outlines of facial components. Different from existing face SR methods, we propose a CSP layer, which is integrated into the generative network G, to assist the process of reconstruction on the basis of semantic guidance. Moreover, a new gradient loss is utilized to constrain the solution space and recover high-frequency textures. Meanwhile, we design a discriminator with an additional branch which classifies the semantic categories of images. Experiments demonstrate that CSPGAN outperforms the state-of-the-art approaches.

However, our CSPGAN still has some limitations on the reconstruction of tiny images, *e.g.*, 32 × 32 sized HR images in Fig. 9. The reasons come from two issues. Firstly, the purpose of designing our SSP maps is to represent the image's semantic prior. They are proved to be essential for generator G. Based on an existing facial landmark detector, SSP maps may crash when facing a tiny image. Secondly, fewer pixel in tiny image makes CSPGAN hard to capture valid information and degrades to common facial SR methods. In future study, we plan to investigate the progressive reconstruction of super-resolved facial images.

REFERENCES

[1] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 115–134, 2007.
[2] X. Chen, H. Wu, X. Jin, and Q. Zhao, "Face illumination manipulation using a single reference image by adaptive layer decomposition," *IEEE Trans. Image Process.*, vol. 22, no. 11, pp. 4249–4259, Nov. 2013.
[3] X. Chen, M. Chen, J. Xin, and Q. Zhao, "Face illumination transfer through edge-preserving filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 281–287.
[4] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.
[5] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau, "Eigenface-domain super-resolution for face recognition," *IEEE Trans. Image Process.*, vol. 12, no. 5, pp. 597–606, May 2003.
[6] S. Serikawa and H. Lu, "Underwater image dehazing using joint trilateral filter," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 41–50, 2014.
[7] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 972–986, May 2012.
[8] W. W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 327–340, Jan. 2012.
[9] Z. Hui, W. Liu, and K.-M. Lam, "A novel correspondence-based face-hallucination method," *Image Vis. Comput.*, vol. 60, pp. 171–184, Apr. 2017.
[10] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 614–630.
[11] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 219–235.
[12] Y. Zhuang, J. Zhang, and F. Wu, "Hallucinating faces: LPH super-resolution and neighbor reconstruction for residue compensation," *Pattern Recognit.*, vol. 40, no. 11, pp. 3178–3194, 2007.

[13] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *Proc. 39th AAAI Conf. Artif. Intell.*, Mar. 2015, pp. 1–7.
[14] Y. Xin and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 318–333.
[15] X. Yu and F. Porikli, "Face hallucination with tiny unaligned images by transformative discriminative neural networks," in *Proc. 31st AAAI Conf. Artif. Intell.*, Feb. 2017, pp. 1–7.
[16] Y. Song, J. Zhang, S. He, L. Bao, and Q. Yang, "Learning to hallucinate face images via component generation and enhancement," 2017, *arXiv:1708.00223*. [Online]. Available: https://arxiv.org/abs/1708.00223
[17] C. Yu, T. Ying, X. Liu, C. Shen, and Y. Jian, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2492–2501.
[18] J. Jiang, Y. Yu, J. Hu, S. Tang, and J. Ma, "Deep CNN denoiser and multi-layer neighbor component embedding for face hallucination," 2018, *arXiv:1806.10726*. [Online]. Available: https://arxiv.org/abs/1806.10726
[19] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[20] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 425–434, Aug. 2005.
[21] A. Chakrabarti, A. N. Rajagopalan, and R. Chellappa, "Super-resolution of face images using kernel PCA-based prior," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 888–892, Jun. 2007.
[22] X. Zhang, S. Peng, and J. Jiang, "An adaptive learning method for face hallucination using locality preserving projections," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–8.
[23] H. Huang, H. He, X. Fan, and J. Zhang, "Super-resolution of human face image using canonical correlation analysis," *Pattern Recognit.*, vol. 43, no. 7, pp. 2532–2543, 2010.
[24] P. Jeong-Seon and L. Seong-Whan, "An example-based face hallucination method for single-frame, low-resolution facial images," *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1806–1816, Oct. 2008.
[25] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun./Jul. 2004, p. I.
[26] X. Ma, J. Zhang, and C. Qi, "Hallucinating face by position-patch," *Pattern Recognit.*, vol. 43, no. 6, pp. 2224–2236, 2010.
[27] J. Jiang, R. Hu, C. Liang, Z. Han, and C. Zhang, "Face image super-resolution through locality-induced support regression," *Signal Process.*, vol. 103, pp. 168–183, Oct. 2014.
[28] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1268–1281, Aug. 2014.
[29] J. Jiang, J. Ma, C. Chen, X. Jiang, and Z. Wang, "Noise robust face image super-resolution through smooth sparse representation," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3991–4002, Nov. 2017.
[30] J. Jiang, C. Chen, K. Huang, Z. Cai, and R. Hu, "Noise robust position-patch based face super-resolution via tikhonov regularized neighbor representation," *Inf. Sci.*, vols. 367–368, pp. 354–372, Nov. 2016.
[31] J. Jiang, Z. Wang, C. Chen, and T. Lu, "L1-L1 norms for face super-resolution with mixed Gaussian-impulse noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2089–2093.
[32] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.
[33] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Advances and challenges in super-resolution," *Int. J. Imag. Syst. Technol.*, vol. 14, no. 2, pp. 47–57, 2010.
[34] T. Köhler, X. Huang, F. Schebesch, A. Aichert, A. Maier, and J. Hornegger, "Robust multiframe super-resolution employing iteratively re-weighted minimization," *IEEE Trans. Comput. Imag.*, vol. 2, no. 1, pp. 42–58, Mar. 2016.
[35] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 25–47, 2000.
[36] W. Gong, L. Hu, J. Li, and W. Li, "Combining sparse representation and local rank constraint for single image super resolution," *Inf. Sci.*, vol. 325, pp. 1–19, Dec. 2015.
[37] J. Jiang, X. Ma, Z. Cai, and R. Hu, "Sparse support regression for image super-resolution," *IEEE Photon. J.*, vol. 7, no. 5, pp. 1–11, Oct. 2015.
[38] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[39] L. Yue, H. Shen, J. Li, Q. Yuanc, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Process.*, vol. 128, pp. 389–408, Nov. 2016.
[40] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 370–378.
[41] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3194–3207, Jul. 2016.
[42] H. Shen, L. Peng, L. Yue, Q. Yuan, and L. Zhang, "Adaptive norm selection for regularized image restoration and super-resolution," *IEEE Trans. Cybern.*, vol. 46, no. 6, pp. 1388–1399, Jun. 2016.
[43] S. Baker and T. Kanade, "Hallucinating faces," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 83–88.
[44] C. Liu, H.-Y. Shum, and C.-S. Zhang, "A two-step approach to hallucinating faces: Global parametric model and local nonparametric model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2001, p. 1.
[45] L. An and B. Bhanu, "Face image super-resolution using 2D CCA," *Signal Process.*, vol. 103, pp. 184–194, Oct. 2014.
[46] H. Huang and N. Wu, "Fast facial image super-resolution via local linear transformations for resource-limited applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 10, pp. 1363–1377, Oct. 2011.
[47] G. Gao and J. Yang, "A novel sparse representation based framework for face image super-resolution," *Neurocomputing*, vol. 134, pp. 92–99, Jun. 2014.
[48] C.-T. Tu and J.'-R. Luo, "Robust face hallucination using ensemble of feature-based regression functions and classifiers," *Image Vis. Comput.*, vol. 44, pp. 59–72, Dec. 2015.
[49] Y. Li and X. Lin, "An improved two-step approach to hallucinating faces," in *Proc. 3rd Int. Conf. Image Graph. (ICIG)*, Dec. 2004, pp. 298–301.
[50] S. S. Rajput, A. Singh, K. V. Arya, and J. Jiang, "Noise robust face hallucination algorithm using local content prior based error shrunk nearest neighbors representation," *Signal Process.*, vol. 147, pp. 233–246, Jun. 2018.
[51] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2016, pp. 1–10.
[52] 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.
[53] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.
[54] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[55] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," 2015, *arXiv:1511.05666*. [Online]. Available: https://arxiv.org/abs/1511.05666
[56] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 391–407.
[57] W. Shi, J. Caballero, and F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[58] S. Gross and M. Wilber. (2016). *Training and Investigating Residual Nets*. [Online]. Available: http://torch.ch/blog/2016/02/04/resnets.html
[59] M. Mathieu, C. Couprie, and Y. Lecun, "Deep multi-scale video prediction beyond mean square error," 2015, *arXiv:1511.05440*. [Online]. Available: https://arxiv.org/abs/1511.05440
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[61] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

LU LIU received the B.S. and M.S. degrees in computer science and technology from Shanxi University, in 2010 and 2013, respectively. She is currently pursuing the Ph.D. degree in signal and information processing with the Institute of Information Science, Beijing Jiaotong University, China. Her research interests include face progression, face recognition, computer vision, and deep learning.

SHENGHUI WANG received the B.S. degree in automatic control from the Department of Communication and Control, Northern Jiaotong University, in 2001, and the Ph.D. degree in signal and information processing from the Institute of Information Science, Beijing Jiaotong University, China, in 2007, where he is currently an Assistant Professor. His research interests include wireless networking, communications signal processing, and embedded systems.

LILI WAN received the Ph.D. degree in computer application from Beihang University, in 2007, with a focus on 3D shape retrieval. From 2014 to 2015, she was a Visiting Researcher with the GrUVi Laboratory, Simon Fraser University (SFU), Canada. She is currently an Associate Professor with the Institute of Information Science, Beijing Jiaotong University, China. Her current research interests include shape analysis, 3D vision, VR, and AR.