IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 6, 2019, accepted July 22, 2019, date of publication August 8, 2019, date of current version September 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2934078

Reference Based Face Super-Resolution

ZHI-SONG LIU, (Student Member, IEEE), WAN-CHI SIU, (Life Fellow, IEEE), AND YUI-LAM CHAN, (Member, IEEE)

Center of Multimedia Signal Processing, Department of Electronic and Information Engineering (EIE), The Hong Kong Polytechnic University, Hong Kong

Corresponding author: Zhi-Song Liu (zhisong.ra.liu@connect.polyu.hk)

This work was supported in part by the Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, through the Internal Research Grant (ZZHR), and in part by the RGC Project of the Hong Kong Special Administrative Region, China, under Grant PolyU 152208/17E.

ABSTRACT Despite the great progress of image super-resolution in recent years, face super-resolution has still much room to explore good visual quality while preserving original facial attributes for larger up-scaling factors. This paper investigates a new research direction in face super-resolution, called Reference based face Super-Resolution (RefSR), in which a reference facial image containing genuine attributes is provided in addition to the low-resolution images for super-resolution. We focus on transferring the key information extracted from reference facial images to the super-resolution process to guarantee the content similarity between the reference and super-resolution image. We propose a novel Conditional Variational AutoEncoder model for this Reference based Face Super-Resolution (RefSR-VAE). By using the encoder to map the reference image to the joint latent space, we can then use the decoder to sample the encoder results to super-resolve low-resolution facial images to generate super-resolution images with good visual quality. We create a benchmark dataset on reference based face super-resolution (RefSR-Face) for general research use, which contains reference images paired with low-resolution images of various pose, emotions, ages and appearance. Both objective and subjective evaluations were conducted, which demonstrate the great potential of using reference images for face super-resolution. By comparing it with state-of-the-art super-resolution approaches, our proposed approach also achieves superior performance.

INDEX TERMS Face super-resolution, deep feature extraction, style transfer.

I. INTRODUCTION

The traditional Single Image Super-Resolution (SISR) is defined as using a single low-resolution (LR) images to recover the corresponding high-resolution (HR) image. It has received substantial attention in image processing and computing vision fields. Inspired by the advanced convolutional neural network in various computing tasks, single image SR gains great improvement recently. Face SR can be regarded as a domain-specific SR application. It also can be called face hallucination. The main goal of SR is to have large up-scaling factors, e.g. $8\times$, for practical application, like digital image display, entertainment and video coding assistance. Generally, we can classify face SR approaches into two categories: conventional learning based face SR and deep learning based face SR.

For conventional learning based face SR, researchers tackle the whole image SR to overlapped patches reconstruction. In the early studies, researchers used internal and external images to explore the patch statistics. In order to learn the mapping relationship between LR and HR patches, there are many proposed techniques utilizing classification algorithms to cluster patches into groups for subspace linear estimation. For example, [1], [2] propose to use k Nearest Neighbor (kNN) to search nearest patches online for patch reconstruction. References [3], [4] propose to learn the coupled overcomplete dictionary for sparse representation of the patch pairs. Furthermore, [6] comes up with multiple layers of kNN super-resolution to gradually improve SR quality. Reference [7] makes use of the local geometrical structure on the HR manifold rather than LR manifold as constraints to learn the couple dictionary for sparse reconstruction. Similar idea is also proposed [8], which combines two different regularizations: kernel Hilbert space constraint and HR local geometrical manifold for better SR reconstruction. Besides these domain specific approaches targeted on facial images, there are also some general image SR approaches that can be directly used for face SR. References [10]–[14] use random forests to process a large number of training data so that various patch patterns can be modeled differently by different

The associate editor coordinating the review of this article and approving it for publication was Senthil Kumar.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

decision trees. By stacking multiple decision trees, we can hierarchically reconstruct the LR images layer by layer until obtaining the optimal SR images.

Recently, deep learning based SR approaches have greatly advanced the state-of-the-art performance of SR. Most existing approaches focus on face SR with large up-scaling factors, i.e., 8×. The evaluation of SR performance is inherited the Mean Squared Errors (MSE) pixel loss used in general image SR. For example, general deep learning based SR approaches [18]–[25] can be directly used in face SR. Due to the lack of study on facial priors, general deep learning based SR approaches give general image patterns features rather than facial features so that they usually fail to generate clear and pleasing face SR. There are also some approaches that use facial prior information to perform post-processing on the SR image to enhance the details. However, due to the ill-posed nature of image SR, most SR approaches suffer from blurry results because the fine details are lost in the LR facial images. Recently, researchers come up with two solutions to obtain face SR with good visual quality: Generative Adversarial Network (GAN) [17] to learn the photo-realistic SR images using adversarial loss and perceptual loss and neural texture transfer [26] to transfer the desired textures of the reference images to the SR images.

GAN based face SR approaches [28]–[35] make use of both discriminators and generators to learn the distribution of the HR images to generate ''fake'' images that cannot be distinguished from the HR images by the discriminator. In order to learn the natural textures, a feature loss is added along with MSE loss to balance the distortion and perception of the image SR. Reference [28] can be considered as the first GAN based face SR that ultra-resolves the LR images by 8×. Reference [31] then proposes the transformative discriminative autoencoders to solve the face SR. There are facial priors can also be embedded into face SR to obtain better performance. Reference [29] uses Bi-Network to learn the dense face correspondence field as the prior to guide face SR. Reference [30], on the other hand, uses the recurrent policy network to focus on different regions of the LR images for local enhancement. Reference [32] designs a network to learn facial landmarks and parsing from LR images and then use the learned priors to super-resolve LR images by components. To keep the facial identity, [33] cascades the face SR and recognition networks together to output SR images with closest identity distance to the ground truth images. Neural texture transfer [26] was the first work that transfers the style of the reference image to the input image to the output image with reference style. It proposes to use both content loss and style loss to train the deep neural network. Lately, [34] proposes to use neural texture transfer for face style transfer. The idea is to choose a reference image containing desired style and then transfer the style to the LR image for better SR. Similarly, [36] proposes a multi-scale neural transfer to improve the texture similarity.

For large up-scaling face SR (e.g. 8×), despite the good visual quality of using GAN or neural texture transfer, the generated ''fake'' features can cause inconsistent patterns and distorted identity. Reference [37] gives a discussion on the ''mode collapse'' and ''unmatched pattern'' problems. From the perspective of practice, the potential application of face SR is face recognition so that preserving true identity of facial images is vital. In this paper, we propose a novel Conditional Variational AutoEncoder model for Reference based Face Super-Resolution (RefSR-VAE). In addition to using a single LR image for SR, we also add a reference image, hence referred to as Reference based Super-Resolution (RefSR), to guide the super-resolution process where LR and the reference images contain the same person, so that we can obtain SR image with both sharp visual quality and remain the ground truth identity. Nowadays, face recognition cannot guarantee 100% or even high accuracy for low quality facial image. By using an available reference images, RefSR can be used to super-resolve low-resolution facial images to assist people for human recognition. In order to extract useful information from the reference image for arbitrary LR images, we propose a conditional Variational AutoEncoder that the encoder learns the joint latent model between LR and reference images and then the decoder extracts latent parameters from the encoder to generate SR images. To the best of our knowledge, this is the first work that resolve the face SR as reference based face SR. With the help of existing reference images, we can use RefSR-VAE network to super-resolve raw LR images of different poses, illumination, ages and other variations with good visual quality, while preserving true identity. Our contributions include the following.

- We firstly propose a Single Image Super-Resolution via conditional Variational AutoEncoder (SISR-VAE) to use single image to reconstruct LR images. Experiments show that proposed SISR-VAE can provide good and robust SR performance in different conditions. This good performance proves that the proposed VAE model can achieve good performance comparable to other state-of-the-art face SR algorithms.
- After showing the potential of VAE for face SR, we further propose a Reference based face SR via conditional Variational AutoEncoder (RefSR-VAE) to resolve face SR with large up-scaling factors. It includes three parts: Ref-HR encoder, LR decoder and VGG feature extractor. The experiments show that RefSR-VAE can greatly outperform all other state-of-the-art face SR algorithms both qualitatively and quantitatively.
- Finally, we will introduce a new Reference based (RefSR-Face) dataset for the SR of face images for training and testing. And then we conducted both quantitative and qualitative evaluations to measure the face SR. By comparing with the state-of-the-art SR algorithms, our proposed approaches can achieve superior performance.

Fig. 1 shows our experimental results for comparison between SISR and RefSR. The SISR results are obtained by using recent state-of-the-art face SR approach FSRNet [35], and RefSR results are obtained by using our proposed

FIGURE 1. Comparison between SISR and RefSR.

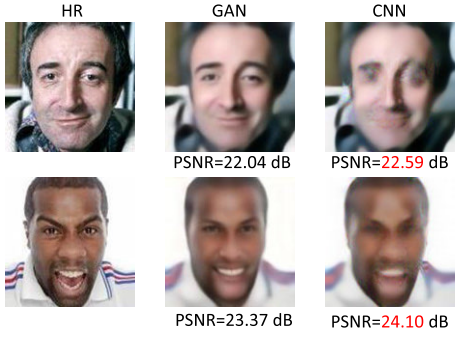

FIGURE 2. Face SR comparison between perception and distortion.

RefSR-VAE. It can be seen that using our proposed approach can achieve superior face SR performance. In Section II, we will give detailed analysis and investigation of face SR. In Section III, we will introduce the proposed RefSR-VAE and RefSR-HAVE in details. In Section IV, we give experimental results to analyze SR performance.

II. RELATED WORK

In this section, let us review the related works from the following perspectives.

A. PHOTO-REALISTIC FACE SR

Photo-realistic face super-resolution focuses on super-resolution with perceptual quality over distortion. The differences between perception (no ground truth reference and quantified by real or fake based on human opinions) and distortion (requires ground truth reference for pixel based measurement) can be explained in Fig. 2.

In Fig. 2, we show two different face SR results: GAN and CNN. GAN represents the SR approach that uses GAN network to minimize perception loss while CNN represents the SR approach that uses the CNN network to minimize the MSE loss. From PSNR results, we can see that results of the CNN approach give lower distortion compared to GAN results. However, from the visual quality, GAN gives sharper reconstruction of facial features. This trade-off between perception and distortion is discussed in [38]. It can be observed from the Maximum A Posterior (MAP) problem of image SR.

$$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}} \log P(\mathbf{X}|\mathbf{Y}) + \log P(\mathbf{Y}) \tag{1}$$

where $\mathbf{Y}$ is the HR image, $\hat{\mathbf{Y}}$ is the estimated SR image, $\mathbf{X}$ is the LR image, logP($\mathbf{X}|\mathbf{Y}$) represents the log-likelihood of LR images given HR images and log P($\mathbf{Y}$) is the prior of HR images that is used for optimization. Formally, we can turn Eq. (1) making use of the MSE minimization,

$$\hat{\mathbf{Y}} = \arg\min_{\mathbf{Y}} \|\mathbf{Y} - \mathbf{C}X\|^2 + \lambda\Omega(\mathbf{Y}) \tag{2}$$

where $\mathbf{C}$ is the mapping model, λ is the weighting parameter and $\Omega(\mathbf{Y})$ is the regularization term. Or we can turn Eq. (1) as divergence minimization,

$$\hat{\mathbf{Y}} = \arg\min_{\mathbf{Y}} d\left[P(\mathbf{Y}) - P(\mathbf{CX})\right] + \lambda\Omega P(\mathbf{Y}) \tag{3}$$

where d is the perceptual loss that measures the divergence between distributions, e.g. the Kullback-Leibler (KL) divergence, Total-Variation (TV) distance, etc. Besides these measurements, GAN based image SR utilizes the discriminator (D) and generator (G) to calculate the adversarial loss to simulate the perceptual loss as

$$\hat{\mathbf{Y}} = \arg\min_{\mathbf{Y}} \sum_{n=1}^{N} \log\left[1 - D_{\theta_D}\left(G_{\theta_G}(\mathbf{X})\right)\right] \tag{4}$$

where N is the batch size. $D_{\theta_D}\left(G_{\theta_G}(\mathbf{X})\right)$ is the probability that the SR image $G_{\theta_G}(\mathbf{X})$ is a natural HR image. It can be interpreted that the generator learns the "fake" sample that cannot be distinguished from its ground truth reference by the discriminator. The discriminator only looks for the meaningful semantic similarity to guide the reconstruction of the generator.

B. REFERENCE BASED IMAGE SUPER-RESOLUTION

Reference based image Super-Resolution (RefSR) (as shown in Fig. 3) is benefited from the study of style transfer [26]. It is an artistic image generation model that can generate the artificial images with the style of the reference image. It uses the covariance of the feature maps of the target and reference images as style reconstruction loss,

$$l^{\phi,j}_{style}(\hat{\mathbf{Y}}, \mathbf{Y}) = \left\| G^{\phi}_{j}(\hat{\mathbf{Y}}) - G^{\phi}_{j}(\mathbf{Y}) \right\|^2_F$$
$$where\ G^{\phi}_{j}(\mathbf{Y})_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H} \sum_{w=1}^{W} \phi_j(\mathbf{Y})_{h,w,c} \phi_j(\mathbf{Y})_{h,w,c'} \tag{5}$$

where $\phi_j(x)$ is the feature map at j-th layer of the network of shape $C_j \times H_j \times W_j$. The Gram matrix $G^{\phi}_{j}(x)$ is calculated as the auto-correlation of $\phi_j(\mathbf{Y})$ with size $C_j \times C_j$. The first example in Fig. 3 shows that the reference style of painting "Candy" is extracted and transferred by a feed-forward network to the target image "Chicago".

Lately, [36] proposes the RefSR approach that utilizes the style reconstruction loss to fuse the reference textures to the

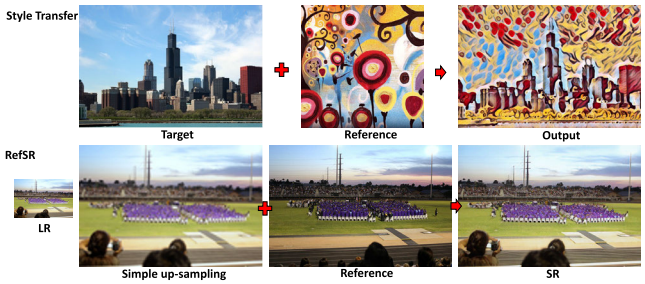

FIGURE 3. Style transfer and RefSR.

LR images to generate SR images with good visual quality. As shown in Fig. 3, using a reference image that contains similar contents super-resolve the LR image to obtain a SR image with better quality. RefSR can use external references found from adjacent frames in a video or the Internet to reconstruct the missing fine details of LR images. The problem is that RefSR still requires the user to find the reference image that is similar to the LR images. The closer the reference image we have, the better SR image we can get. RefSR has another problem that for the same network, each reference requires one training process. During testing, it means that for each LR image we not only need to find the similar reference image, we also have to train the network from scratch. This online training process is very time consuming for real-time application.

Face SR as a domain specific application of image SR, it has one advantage that the faces share similar statistics. Considering the potential application of face SR, each identity can have many facial images captured at different conditions and stored in the dictionary. We can use RefSR for face SR to improve SR performance based on the following procedure: for each identity, we store one HR facial image that captures the frontal face with clear facial features (no need for alignment) as the Reference. Then we have LR and HR image pairs of the same person captured at different conditions. We form tripled Ref-LR-HR images for each identity and train a model for super-resolution. With the help of reference images, the target is to train one model for face SR across different identity at different conditions so that we can avoid online training process. During testing, we only need to input LR images along with their corresponding reference images to the model, and then we can generate the SR image. In the following sections, we will introduce the propose RefSR-VAE model.

III. THE PROPOSED WORK

In this sections, we will introduce the proposed RefSR-VAE. As shown in Fig. 4, the proposed RefSR-VAE contains three parts: i) Ref-HR encoder which learns the conditional generative model (approximates the posterior distribution of the latent variables) between the reference and LR images, ii) SR decoder which reconstructs the SR images from sampled latent variables, and iii) VGG feature extractor which calculates the perceptual loss using pretrained networks. The proposed structure is built upon the idea of the conditional variational autoencoder. We will first give a brief review of the conditional variational autoencoder. Then we will introduce the propose RefSR-VAE. Finally, we will introduce the new Reference based SR of face image (RefSR-Face) dataset for training and testing.

A. CONDITIONAL VARIATIONAL AUTOENCODER

Let us formally introduce the Conditional Variational Autoencoders for image super-resolution. We denote the LR image by $\mathbf{X}\in \mathbb{R}^{m\times n\times 3}$, and the HR image by $\mathbf{Y}\in \mathbb{R}^{\alpha m\times \alpha n\times 3}$, where α is the up-sampling factor and m, n is the dimension of the image. Given a vector of z in a high-dimensional space $\mathbf{Z}$, the goal of the conditional variational autoencoders is to learn the conditional generative model as,

$$P(\mathbf{Y}|\mathbf{X}) = \int P(\mathbf{Y}|\mathbf{X}, z)P(z|\mathbf{X})dz \tag{6}$$

Generally, given arbitrary z sampled from some distribution, it is almost not possible to obtain desired posterior $P(\mathbf{Y}|\mathbf{X})$. We want to arrange the network to learn parameters θ to maximize the data log likelihood $P_\theta(\mathbf{Y}|\mathbf{X})$ as,

$$\log P_\theta(Y|X) \geq \int \log P(Y|X, z)P(z|X)dz \tag{7}$$

we can use Bayesian rule to rewrite Eq. (7) as

$$\begin{aligned}\log P_\theta(\mathbf{Y}|\mathbf{X}) &\geq \int \log P(\mathbf{Y}|\mathbf{X}, z)P(z|\mathbf{X})dz \\ &= E_{Q_\phi(z|\mathbf{X})}\left[\log \frac{P_\theta(\mathbf{Y}|\mathbf{X}, z)P_\theta(z|\mathbf{X})}{Q_\phi(z|\mathbf{X}, \mathbf{Y})}\right]\end{aligned} \tag{8}$$

Note that the key of conditional variational autoencoders is to learn the latent vector z first and then we sample from z to find the posterior $P(\mathbf{Y}|\mathbf{X})$. We can use KL to represent the divergence between predicted distributions $P_\theta(z|\mathbf{X})$ and $Q_\phi(z|\mathbf{X},\mathbf{Y})$. We have the following equation,

$$\begin{aligned}\log P_\theta(\mathbf{Y}|\mathbf{X}) &= E_{Q_\phi(z|\mathbf{X})}\left[\log \frac{P_\theta(\mathbf{Y}|\mathbf{X}, z)P_\theta(z|\mathbf{X})}{Q_\phi(z|\mathbf{X}, \mathbf{Y})}\right] \\ &= E_{Q_\phi(z|\mathbf{X})}\left[\log P_\theta(\mathbf{Y}|\mathbf{X}, z)\right] \\ &\quad -KL\left[Q_\phi(z|\mathbf{X}, \mathbf{Y})|P_\theta(z|\mathbf{X})\right]\end{aligned} \tag{9}$$

where KL[$p|q$] represents the KL divergence. Eq. (9) can be interpreted in the way that we use encoder to learn an approximation to the posterior $Q_\phi(z|\mathbf{X}, \mathbf{Y})$ and then use decoder to learn the tractable likelihood $P_\theta(\mathbf{Y}|\mathbf{X}, z)$. To apply the mini-batch gradient optimization, we use the "reparameterization trick" proposed in [40] and to use $\varepsilon \sim \mathrm{N}(0, I)$ to randomly sample from $Q_\phi(z|\mathbf{X}, \mathbf{Y})$ and then compute $z = \mu(\mathbf{X},\mathbf{Y})+\Sigma^{0.5}(\mathbf{X},\mathbf{Y})*\varepsilon$. Hence, we can calculate the gradient of Eq. (9) as,

$$\begin{aligned}&\frac{1}{N}\sum_n \log P_\theta(\mathbf{Y}_n|\mathbf{X}_n) \\ &= \frac{1}{N}\sum_n E_{Q_\phi(z|\mathbf{X})}\end{aligned}$$

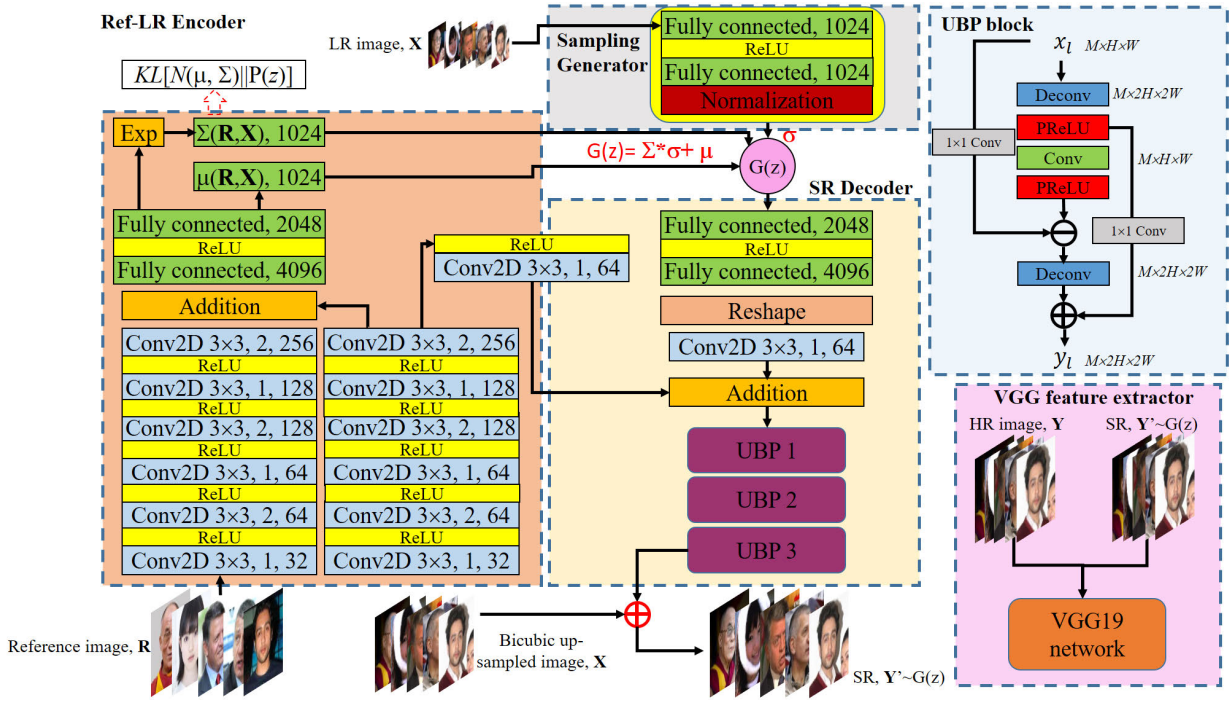

FIGURE 4. Complete structure of the proposed RefSR-VAE.

$$\left[\log P_\theta(\mathbf{Y}|\mathbf{X}, z = \mu(\mathbf{X}_n, \mathbf{Y}_n) + \Sigma^{0.5}(\mathbf{X}_n, \mathbf{Y}_n)^*\varepsilon)\right] - KL\left[Q_\phi(z|\mathbf{X}_n, \mathbf{Y}_n)|P_\theta(z|\mathbf{X}_n)\right] \quad (10)$$

We assume that $P_\theta(z|\mathbf{X})$ is independent from training data $\mathbf{X}$ so we randomly sample z from $N(0, I)$ for training. Hence, at test time, we simply input $z \sim N(0, I)$ into the decoder to generate the new sample.

B. REFERENCE BASED FACE SR VIA CONDITIONAL VARIATIONAL AUTOENCODER

To utilize conditional variational autoencoder for Reference based face SR, we only need to change posterior from $P(\mathbf{Y}|\mathbf{X})$ to $P(\mathbf{R}|\mathbf{X})$, where $\mathbf{R}$ is reference facial image of the same identity. Then we can rewrite Eq. (10) as,

$$\frac{1}{N}\sum_n \log P_\theta(\mathbf{Y}_n|\mathbf{R}_n, \mathbf{X}_n) = \frac{1}{N}\sum_n E_{Q_\phi(z|\mathbf{RX})} \left[\log P_\theta(\mathbf{R}|\mathbf{X}, z = \mu(\mathbf{R}_n, \mathbf{Y}_n) + \Sigma^{0.5}(\mathbf{R}_n, \mathbf{Y}_n)^*\varepsilon)\right] - KL\left[Q_\phi(z|\mathbf{R}_n, \mathbf{X}_n)|P_\theta(z|\mathbf{R}_n, \mathbf{X}_n)\right] \quad (11)$$

The target is to use encoder to learn the latent relationship between LR and reference images $P(z|\mathbf{R}, \mathbf{X})$. Hence, for each identity, we map various LR facial images taken under different conditions with the reference image to extract the most representative attributes that are robust against different conditions. This rigid mapping relationship is based on the assumption that the images of the same person share the same facial attributes so that this one-to-many correlation can be converted to the latent space and sample a latent vector for super-resolution. In order to perform super-resolution to the LR images, the decoder can sample from the learned generative model of $P(z|\mathbf{R}, \mathbf{X})$ to generate SR images.

The complete structure of the proposed RefSR-VAE is shown in Fig. 4. The model includes four parts: Ref-LR Encoder, Sampling Generator, SR Decoder and VGG feature extractor. Ref-LR Encoder works as a generative model that learns the latent variables z of the distribution of $P(z|\mathbf{R},\mathbf{X})$. The input is paired Bicubic up-sampled LR images and their reference images. The output is the approximation of mean $\mu(\mathbf{R},\mathbf{X})$. and variance $\Sigma(\mathbf{R},\mathbf{X})$ of training data $Q_\phi(z|\mathbf{R},\mathbf{X})$. To reduce the KL divergence between $Q_\phi(z|\mathbf{R},\mathbf{X})$ and $P(z|\mathbf{R},\mathbf{X})$, we assume that $P(z|\mathbf{R},\mathbf{X})$ follows the Gaussian distribution $P(z) \sim N(\mu_N, \Sigma_N)$ that is independent of the training data. The computation of KL is,

$$L_{KL} = \log\frac{\sum_N}{\sum_Q} + \frac{\sum_Q + (\mu_Q - \mu_N)^2}{2\sum_p} \quad (12)$$

Usually, we set $\mu_N = 0$ for simplicity. Reference [41] discusses about the choice of Σ_N for computing KL. A smaller Σ_N suggests a narrower searching space of z that can generate sharper samples but bizarre pattern while a larger Σ_N generates blurry and plausible samples. To encourage better SR results, we set $\sigma = 0.1$ experimentally.

The second part is the Sampling generator. It works as sampling process of "reparameterization trick" for training. In most of VAE based works [39]–[41], they use simple Gaussian distribution $\varepsilon \sim N(0, I)$ (as introduced in the previous

subsection) to complete the training procedure as described in Eq. (10). We propose to use sampling generator G(z) to learn the subspace of latent variables for sampling. Details of sampling generator G(z) are shown in Fig. 4. We design two layers of fully connected layers followed by a normalization layer to normalize the output to mean 0 and variance σ as the sampling distribution. During training, instead of using $\varepsilon \sim N(0, I)$, we use G(z) to randomly sample from "encoder" distribution as $\mathrm{G}(z) = \Sigma^{*}\sigma + \mu$ (latent variables learned from the Ref-LR images) and the SR decoder uses the sampled distribution to reconstruct the SR image. After training, we can directly sample from normal distribution $P(z) \sim N(\mu_N, \Sigma_N)$ to output the SR results.

Next, we arrange the SR Decoder to learn the reconstruction process of SR. The basic module is Up-sampling Back Projection (UBP) block. It is based on previous studies of back projection based residual network on image SR [23]–[25]. Inside the UBP block, as shown in the right up corner of Fig. 4, it is designed based on the concept of back projection: to improve data fidelity of SR, we minimize the loss between the original LR image and the down-sampled SR image. The same idea used in UBP can be described mathematically as,

$$y_l = W_1 f(U \otimes x_l) + U \otimes (W_2 x_l - f(D \otimes f(U \otimes x_l))) \qquad (13)$$

where x_l is the input (LR) feature maps of size $M \times H \times W$. y_l is the output (HR) feature maps of size $M \times 2H \times 2W$. M is the number of feature maps and H and W are the sizes of the feature map. W_1 and W_2 are the weighting processing that use 1×1 convolutional layer. U and D are the convolutional and deconvolutional layers that work as the up-sampling and down-sampling processes. Each UBP block achieves $2\times$ up-sampling. Depends on different up-sampling factors, we can stack different number of UBP blocks to achieve the goal. In Fig. 4, we show the case for $8\times$ face SR so that we stack three UBP blocks. After the final UBP block, we add the shortcut form the input Bicubic up-sampled LR image to form the output of UBP blocks to obtain the final SR images $\mathbf{Y}$'. To guarantee the data fidelity, we calculate the pixel based Mean Absolute Errors (MAE) as follows,

$$L_{MAE} = \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} |\mathbf{Y}_{c,h,w} - \mathbf{Y}'_{c,h,w}| \qquad (14)$$

where C, H and W are the size of HR images.

Similar to GAN base image SR, to encourage the network to generate photo-realistic images, we also arrange the VGG feature extractor to learn the semantic similarity between SR and HR images using distance in the deep feature space. We can use pre-trained VGG19 [16] to extract the feature maps for estimation. Both SR and HR images are sent to VGG19 (fix the parameters) which outputs the corresponding feature maps obtained by the 4th convolution layer before the 5th "Maxpooling" layer. We define SR and HR feature maps as $\Phi_{54}(\mathbf{Y}')$ and $\Phi_{54}(\mathbf{Y})$. Inspired by [27], we suggest to extract the feature maps before the "ReLU" activation layer

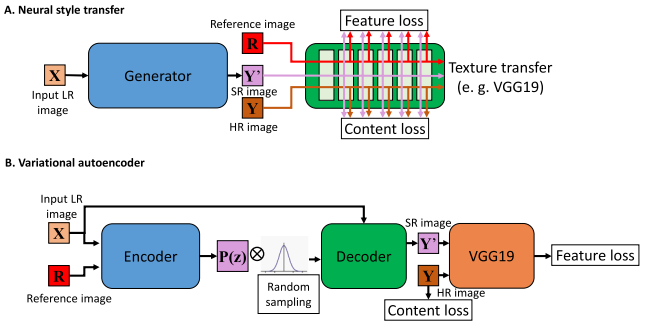

FIGURE 5. Comparison between neural style transfer and conditional variational autoencoder for RefSR.

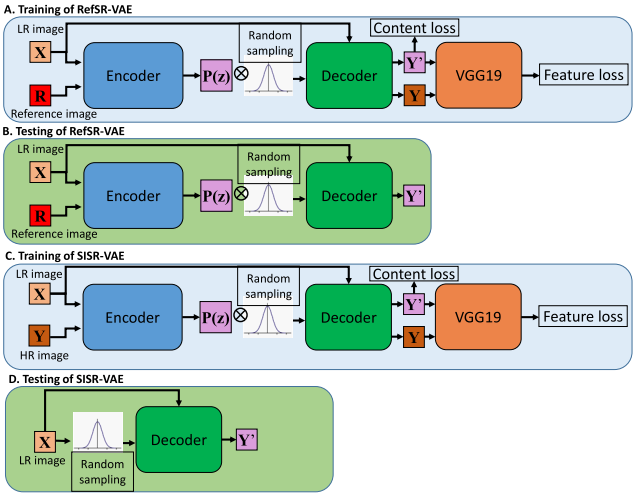

FIGURE 6. Training and testing process of SISR-VAE and RefSR-VAE.

because we want to use all feature response to calculate the feature loss as,

$$L_{VGG} = \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} |\Phi_{54}^{c,h,w}(\mathbf{Y}) - \Phi_{54}^{c,h,w}(\mathbf{Y}')| \qquad (15)$$

Finally, we have the total loss of combination of pixel based MAE loss L_{MAE}, VGG feature loss L_{VGG} and also KL loss L_{KL} (Eq. 12) to average the mini-batch gradients for backpropagation.

C. COMPARISON BETWEEN VARIATIONAL AUTOENCODERS AND NEURAL STYLE TRANSFER

After introducing RefSR-VAE, let us discuss the differences between neural style transfer and our proposed variational autoencoders for reference image SR. Let us use Fig. 5 to simplify the structure of two different models.

The neural style transfer is built based on a feed-forward network to transfer the style of the reference image to the input LR image to aid the super-resolution process. As shown in Fig. 5A, the neural style transfer usually includes two subnetworks: generator and texture transfer. The generator super-resolves LR images $\mathbf{X}$ to generate SR images $\mathbf{Y}$'. In order to improve the SR quality, we need to find an image with contents similar to LR image as the reference image $\mathbf{R}$.

FIGURE 7. Samples of RefSR-Face dataset.

Then there is the texture transfer that calculates the feature loss between R and **Y**' using the style reconstruction loss (introduced in Eq. (5)) and the content loss between **Y**' and **Y**. The content loss can be regarded loosely as a per-pixel loss calculated at some extracted feature maps of the texture transfer. Similarly, the style reconstruction loss can also be calculated at extracted feature maps of the texture transfer.

For variational autoencoder, it contains an encoder, a decoder and a VGG feature extractor. The encoder learns the latent variable model of the correlation between reference and LR images. Then the decoder samples from the encoded latent space to super-resolve LR images. The losses include three parts: KL divergence, content loss and feature loss.

Both neural style transfer and variational autoencoder can be used for face SR. In essence, the former is still a discriminative model to maximize posterior likelihood of LR super-resolution, while the latter is a generative model to capture the dependencies between reference and LR images. The disadvantage of neural style transfer is that it requires online training. Given a LR image, it needs to collect reference patches across different scales to perform patch matching and feature swapping. It is time consuming for real-time application. For variational autoencoder, it has encoded the dependencies between LR and its reference images during training. During testing, we only need to sample from the learned distribution for reconstruction.

Finally, not only we can use variational autoencoder for RefSR, we can also use it for Single Image Super-Resolution (SISR) which only uses LR images for SR. We call it SISR-VAE in the following discussion.

The training and testing process of SISR-VAE and RefSR-VAE are shown in Fig. 6. Both SISR-VAE and RefSR-VAE can make use of the same structure of variational autoencoder introduced in Fig. 4. The only difference is that the encoder of SISR-VAE learns latent distribution of posterior $P(\mathbf{Y}|\mathbf{X})$. During testing, SISR-VAE only needs the decoder to perform super-resolution.

D. REFERENCE BASED FACE SUPER-RESOLUTION DATASET

Finally, we propose a dataset for face RefSR training and testing. The target is to collect facial images across different sex, races and so on. Each identity should include several facial images with various poses, ages, emotions and so on. In [42], authors of the paper proposed a VGGFace2 dataset that contains images downloaded from Google Image Search which has a large variation in pose, age, illumination, ethnicity and profession. The testing dataset includes 500 identities of annotated facial images with different variations.

To form the RefSR dataset, we split VGGFace2 testing dataset to form the RefSR-Face training and testing dataset. The process includes three steps: 1) for each identity, we can select at least two images with size no smaller than 128*128, 2) we crop the image and resize the images to 128×128 by Bicubic using MATLAB and 3) we can manually select the most representative images that contain frontal face for each identity as the reference. Finally, we can then obtain a training dataset containing 428 identities for development. Each identity includes 2∼30 images. And a testing dataset also contains 428 identities. Each identity includes 1∼4 images with very different appearance to the reference image for evaluation. Totally, the testing dataset has 560 images obtain the training dataset has 7451 images.

To summarize this part of the discussion, let us show several cases in RefSR-Face in Fig. 7. For each identity there is a reference image and their HR images. The chosen reference images do not need to be frontal faces, like group D, G and K. For the same person, the chosen reference image can be very different from other HR images. For example, we can have dressing difference (A shows a woman with a hat and J shows a woman with makeup), age difference (C shows childhood of the woman and H shows a woman at different ages) and difference in movie or TV pictures (F shows a black-and-white film shot and L shows another movie shot). Using RefSR-Face dataset, we can evaluate the performance of face RefSR.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

Datasets: We conducted extensive experiments on three available datasets: Helen [43], CelebA [44] and our proposed RefSR-Face. We coarsely cropped the images according to their face regions and resized to 128×128×3 without any pre-alignment operation. Among all face SR approaches, Helen dataset is commonly used for evaluation. We followed the same procedure for comparison. We selected 2330 images for training and the rest 50 images for testing. We call the Helen testing images as Helen-50 for clarity. CelebA dataset is another popular data for comparison. It is a larger dataset that contains facial images with more variations. We used the first 18,000 images for training and the rest 1000 images for testing. We call the CelebA testing images as CelebA-1000. For RefSR-Face, we used 7451 images for training and the rest 560 images for testing. Depending on whether using reference images or not, RefSR-Face has also been used for SISR and RefSR evaluation.

We have only focused on 8× up-sampling. During the training stage, the LR images were down-sampled from HR

images by using *Bicubic* in MATLAB and then we used *Bicubic* again as an initial up-sampling operation to up-sample LR images to the same size as the HR images. The initial up-sampled LR and HR images were formed as image pairs for training. The same as most image SR approaches, we enlarged the training images by image augmentation, including flipping and rotation. Eventually, we were able to generate around 100,000 training data.

Most of our experiments were conducted for 8× face SR. Hence, we used 3 UBP blocks in SR Decoder for three times of 2× up-sampling. Inside the UBP block, we used Parametric ReLU (PReLU) for activation and 6×6 filter with stride 2 for convolution and deconvolution. Pre-trained VGG19 is provided by [16]. For implementation, we trained our model with learning rate 0.0001 for all layers for a total of 500,000 iterations. For optimization, we used Adam with momentum to 0.9 and weight decay 0.0001. All experiments were conducted using Caffe, MATLAB R2016b on a NVIDIA GTX 1080Ti GPU.

Different settings and SR algorithms were evaluated in terms of PSNR and SSIM. They are standard distortion based SR estimation methods which describe the pixel based loss. The same as most face SR algorithms, we converted RGB images to YUV images and only used Y channel for calculation. We ignored 8 pixels at each boundary to avoid the boundary effect.

B. ANALYSIS OF SINGLE IMAGE SUPER-RESOLUTION

Since RefSR for face SR is a new topic in computing vision field, to the best of our investigation, there is no related approaches for comparison. Let us firstly see the efficiency of our proposed SISR-VAE introduced in the previous section. We only have the LR images for SR. In our investigation, we compared SISR-VAE with many state-of-the-art face SR algorithms, including Local Linear Embedding (LLE) [2], Locality constrained Representation (LcR) [7], Super-Resolution via Convolutional Neural Network (SRCNN) [18], Very Deep convolution network for image Super-Resolution(VDSR) [19], Super-Resolution via ResNet (SRResNet) [21], Global Local face SR Network (GLN) [30], Ultra-Resolving face images by Discriminative Generative Networks (UR-DGN) [28] and the recent superior approach Face Super-Resolution with Facial Priors (FSRNet) [35]. Note also that results of SRResNet were provided by the authors of FSRNet [35]. Except this, all other approaches were reimplemented based on the codes provided by the respective authors.

TABLE 1 shows the SR results on Helen-50 and CelebA-1000 datasets. We can see that our SISR-VAE approach can achieve comparable PSNR and SSIM similar to FSRNet, and much higher performance compared to other state-of-the-art algorithms. Note that FSRGAN and FSRNet are the results from [35]. FSRGAN is the GAN version of FSRNet that was trained based on perception loss rather than pixel based MSE loss. FSRGAN can generate images with better visual quality but sacrificing data fidelity.

TABLE 1. State-of-the-art SR algorithms comparison on PSNR and SSIM (red indicates the best and blue is the second best results).

Dataset	Helen-50		CelebA-1000	
Eval.	PSNR	SSIM	PSNR	SSIM
Bicubic	23.69	0.659	23.75	0.642
SRCNN (2016)	23.97	0.678	24.26	0.663
VDSR (2016)	24.61	0.698	24.83	0.688
SRResNet (2017)	25.30	0.730	25.82	0.737
GLN (2016)	24.11	0.692	24.55	0.687
UR-DGN (2016)	24.22	0.691	24.63	0.685
FSRNet (2018)	26.21	0.772	26.60	0.763
FSRGAN (2018)	25.10	0.723	25.20	0.702
SISR-VAE	26.10	0.771	25.86	0.751

In general, our SISR-VAE can achieve the second best performance among different SR algorithms. It is understandable because SRCNN, VDSR and SRResNet were trained on general image set for general image SR. They may not perform well on face SR. GLN and UR-DGN can perform well but the results are not as good as those in FSRNet or SISR-VAE because they can be considered as early investigation of face SR using deep learning approaches. The most competitive approaches are FSRNet and FSRGAN that were proposed recently. In the following discussion, we try to focus more on them for comparison. It is noticeable that our SISR-VAE can achieve similar PSNR and SSIM performance as FSRNet on Helen-50 but it is not very impressive on CelebA-1000. Our explanation is that FSRNet (and FSRGAN) was proposed based on using facial prior information to guide face SR. That is, the model is trained by using both LR images and their priors. The priors include facial parsing and landmark. The authors first prepared these priors along with the facial images and then trained the model to minimizing both pixel based MSE loss and prior losses. Different from Helen dataset, CelebA dataset does not have ground truth parsing maps so the authors use GFC [45] to estimate the parsing map as the pseudo ground truth for training. In order to obtain the optimal performance, the authors separately used Helen and CelebA datasets to train two different sets of models to test on corresponding datasets. Hence, it could be the reason why FSRNet can perform even better than Helen-50 for CelebA-1000.

TABLE 2. FSRNET and SISR-VAE comparison on RefSR-face dataset (red indicates the best and blue is the second best results).

SR algorithms	PSNR	SSIM
Bicubic	22.26	0.569
FSRNet	24.13	0.662
FSRGAN	23.50	0.629
SISR-VAE(independent)	24.18	0.663
SISR-VAE(dependent)	25.82	0.736

In order to verify the general performance of our proposed SIS-VAE. We conducted further experiments on the performance on our proposed RefSR-Face dataset for SISR. Note that RefSR-Face is very different from Helen and CelebA because it contains images with more difficult conditions. It can be used to verify the generalization ability of different face SR approaches. To make the comparison, we call the SISR-VAE in TABLE 1 as SISR-VAE(independent). It trained on the same datasets (Helen and CelebA) as FSRNet and FSRGAN. We also trained another SISR-VAE(dependent) model that used RefSR-Face training set for comparison. We have the results as shown in TABLE 2.

From TABLE 2, we can see that SISR-VAE(dependent) achieves the best PSNR and SSIM. Compared to FSRNet, it can improve nearly 3 dB PSNR and 0.1 SSIM. Even SISR-VAE(independent) can also outperform FSRNet about 0.5 dB in PSNR and 0.02 in SSIM. Comparing SISR-VAE(dependent) and SISR-VAE(independent), it is obvious that using RefSR-Face for training can achieve better performance.

To further demonstrate the superior performance of our proposed SISR-VAE, let us show some SR images for visual comparison. In Fig. 8, we show SR results of Bicubic, FSRNet, FSRGAN and SISR-VAE on Helen-50 and RefSR-Face datasets. It can be see that using proposed SISR-VAE can also achieve good visual quality comparable to FSRNet or FSRGAN. Though FSRGAN can output much sharper SR images compared to other approaches, it also predicts wrong facial features that affect the data fidelity. For example, the left eye of the man in the third row should be close but FSRGAN predicts an open eye. FSRGAN also predicts wrongly about the man in the fourth row that he does not wear glass and has open eyes.

Furthermore, we also show several cases in RefSR-Face to demonstrate the generality of proposed SISR-VAE over FSRNet and FSRGAN.

From Fig. 9, we demonstrate different SR algorithms on RefSR-Face testing set. It can be found that using FSRNet and FSRGAN fail to provide good face SR. The facial features are distorted. Compared to the good SR results on Helen-50 in Fig. 8, the reason that FSRGAN and FSRNet fail to perform on RefSR-Face can be twofold: 1) RefSR-Face contains more challenging facial images that Helen and CelebA do not include. For instance, the last row of Fig. 9,

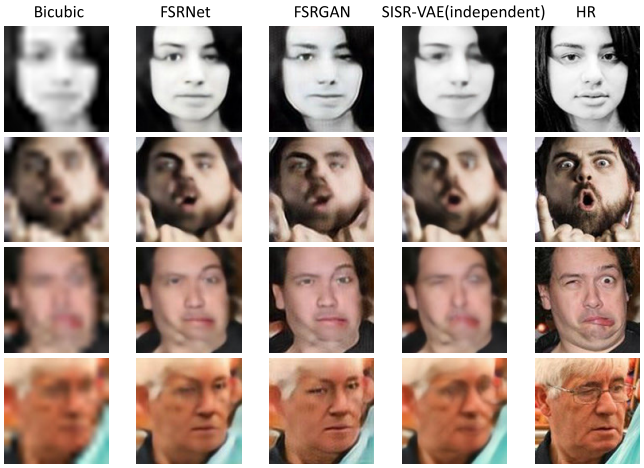

FIGURE 8. Visual comparison among different SR algorithms on Helen-50 testing set.

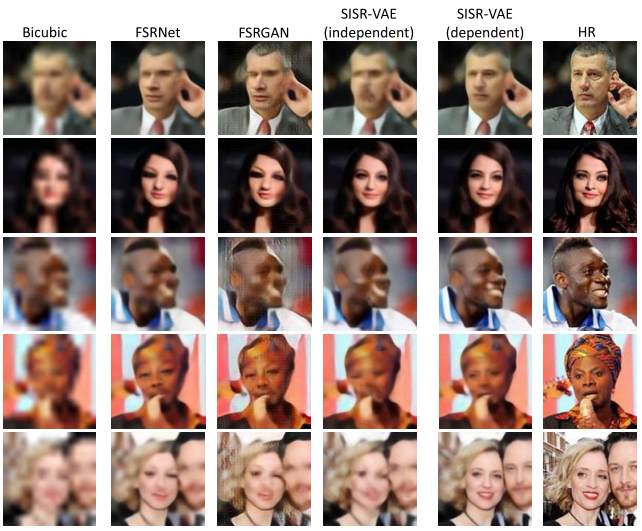

FIGURE 9. Visual comparison among different SR algorithms on RefSR-Face testing set.

the image contains a lady with a partial face of a man; and 2), FSRNet and FSRGAN trained on Helen and CelebA datasets with corresponding facial priors. It could be easily to encounter overfitting problem so that they can perform well on facial images captured in the similar conditions. On the other hand, using SISR-VAE(independent) can still generate SR images with reasonable facial features. The SR visual quality are more consistent from Helen-50 to RefSR-Face. Furthermore, we also have SISR-VAE(dependent) trained on RefSR-Face training set that performs better on the testing set.

C. ANALYSIS OF REFERENCE BASED IMAGE SUPER-RESOLUTION

After illustrating the efficiency of the proposed SISR-VAE, we can formally demonstrate the performance of proposed RefSR-VAE.

FIGURE 10. Visual comparison among different SR algorithms.

RefSR-Face was used to train our RefSR-VAE model introduced in Fig. 4. During the training stage, we used image triplets Ref-LR-HR to train the model. That is, for the same person, we have one reference image and many different HR images and their corresponding LR images (Reference images share the same identities with different HR images but with very different conditions). The Ref-LR Encoder explores the hidden latent variables across different Ref-LR image pairs and the SR Decoder reconstructs LR images based on the guidance of reference images to predict the corresponding HR images. During testing, we input LR images and their corresponding reference images for predicting the SR images. We have obtained the SR performance in the following table.

In TABLE 3, we tested both SISR-VAE and RefSR-VAE on RefSR-Face dataset. To fully explore the potential of RefSR-VAE, we tested RefSR-VAE on three different conditions. RefSR-VAE(reference) is the RefSR results that uses the corresponding reference images to super-resolve LR images. RefSR-VAE(Gaussian) is the RefSR results that uses the random Gaussian noise with 0 mean and 1 variance as reference for SR. RefSR-VAE(LR) is the RefSR results that uses LR images themselves as references for SR. RefSR-VAE(reference) is the best results that used reference images to aid LR facial image super-resolution. Compared to SRCNN, FSRNet and FSRGAN, it can achieve at least 3dB improvement in PSNR and 0.15 improvement in SSIM. RefSR-VAE(reference) can also outperform SISR-VAE about 0.15dB in PSNR and 0.04 in SSIM. RefSR-VAE(LR) and RefSR-VAE(Gaussian) are alternative choices when reference images are missing or we cannot ensure the true identity of the LR facial images. They do not provide good quantitative performance but they are still worthy further investigation. Note that we only used RefSR-Face training set to train RefSR-VAE(reference) model. For testing, depending upon different reference inputs (LR, Gaussian noise and Reference images), we used the RefSR-VAE testing model (as shown in Fig. 6B) to obtain corresponding results of RefSR-VAE(LR), RefSR-VAE(Gaussian) and RefSR-VAE(reference). We show the visualization of different results in Fig. 10 to further demonstrate the effectiveness of the proposed methods.

From Fig. 10, we show the results on facial images with different conditions, including a poster picture (5^{th} row), makeup or movie picture (1^{st} and last row), black&white picture (3^{rd} row) and so on. The reference images are also very different from LR images to avoid any spatial similarity. Compared with our proposed algorithms, both FSRNet and FSRGAN fail to recover fine facial features because the facial

TABLE 3. State-of-the-art SR algorithm comparison on RefSR-face dataset.

SR algorithms	PSNR	SSIM
Bicubic	22.26	0.569
SRCNN	22.57	0.588
FSRNet	24.13	0.662
FSRGAN	23.50	0.629
SISR-VAE(independent)	24.18	0.663
SISR-VAE(dependent)	25.82	0.736
RefSR-VAE(Gaussian)	22.26	0.567
RefSR-VAE(LR)	22.69	0.630
RefSR-VAE(reference)	27.21	0.779

TABLE 4. PSNR and SSIM comparison among different face SR algorithms.

SR algorithms	PSNR	SSIM
Noise		
SRCNN	20.22	0.501
FSRNet	18.85	0.488
FSRGAN	19.24	0.466
SISR-VAE	23.29	0.538
RefSR-VAE	24.25	0.589
Occlusion		
SRCNN	16.12	0.411
FSRNet	14.46	0.398
FSRGAN	14.16	0.322
SISR-VAE	17.14	0.454
RefSR-VAE	19.41	0.495

images have different poses from the training images they used in CelebA and Helen. It is interesting to find that RefSR-VAE(LR) can also be able to reconstruct facial features well even without the aid of reference images. It can be explained that Ref-LR encoder is originally designed to extract the compact model distribution of Ref-LR. When we input LR images as reference, the input becomes LR-LR images with same spatial correlation so that the LR facial edges are multiplied by 2 time for reconstruction. From Fig. 10, we can also observe that strong edges in LR images are further enhanced without much distortion. However, compared to RefSR-VAE(reference), fine details of facial features are still missing.

D. ANALYSIS OF ROBUSTNESS OF FACE SUPER-RESOLUTION

In practical applications, face SR needs to be robust enough against various distortion and interference. A good face SR model should be able to generate SR images with good quality. To test the robustness of face SR, we choose two ways of distortions to conduct the experiments: noise and occlusion.

For image denoising, as studied in many research works [46]–[47], researchers added white Gaussian noise with different variances to simulate the real situation. Similarly, we added random Gaussian noise $N(0,0.05)$ to the LR facial images, and then used the noisy LR images as inputs to different face SR models to perform super-resolution. For image occlusion, we randomly cropped out a 64×64 region from LR facial images and filled it with random Gaussian noise $N(0,1)$. These occluded LR images were then applied to different face SR models to perform super-resolution.

We first calculated the quantitative results as shown TABLE 4. From the results of the table, it can see that using proposed SISR-VAE and RefSR-VAE can achieve better PSNR and SSIM on both denoising and occlusion. In Fig. 11, we show two cases of SR results to show the visual differences.

In Fig. 11, the first row shows the SR results generated from original LR images obtained by different approaches. The second and third rows are the SR results generated from occluded and noisy LR images. Since all the approaches are proposed to super-resolve LR images, if there are occlusions in the images, they all fail to predict the occlusions. Still, the advantages of our proposed SISR-VAE and RefSR-VAE are not affected by the occlusions. The unoccluded parts are still able to be reconstructed. On the other hand, FSRNet and FSRGAN are affected by the existence of occlusions. The unoccluded parts also become blurry and unclear. For instance, image B of Fig. 11, the eyes can be preserved in SISR-VAE and RefSR-VAE but distorted by FSRNet and FSRGAN. For noisy situation, after adding the Gaussian noise, both FSRNet and FSRGAN have global impacts that the details cannot be reconstructed and even the color is distorted. Our proposed approaches are still able to resist the random noise and generate SR images with better visual quality.

Furthermore, we also evaluated different approaches using Labeled Faces in the Wild (LFW) [48] dataset. We choose it because the images were collected from the web. They represent people in a wide variety of settings, poses, expressions and lighting so it can be used for unconstrained face recognition. To perform face SR on this dataset, we randomly chose 200 images from the deep-funneled LFW dataset, for which the images were roughly aligned by deep neural network. We used *Bicubic* function in MATLAB to generate LR images for estimation. We have the SR results in TABLE 5.

From TABLE 5, we can find that SRCNN, FSRNet and FSRGAN fail to generate SR images with good PSNR and SSIM. Since the LFW dataset does not have reference image for each identity, we cannot perform RefSR-VAE for face SR, we only used SISR-VAE to perform face SR. Generally, SISR-VAE outperforms other approaches at least 3 dB in PSNR and 0.12 in SSIM. It demonstrates the generalizability of our proposed SISR-VAE. It can achieve robust SR performance on different facial images.

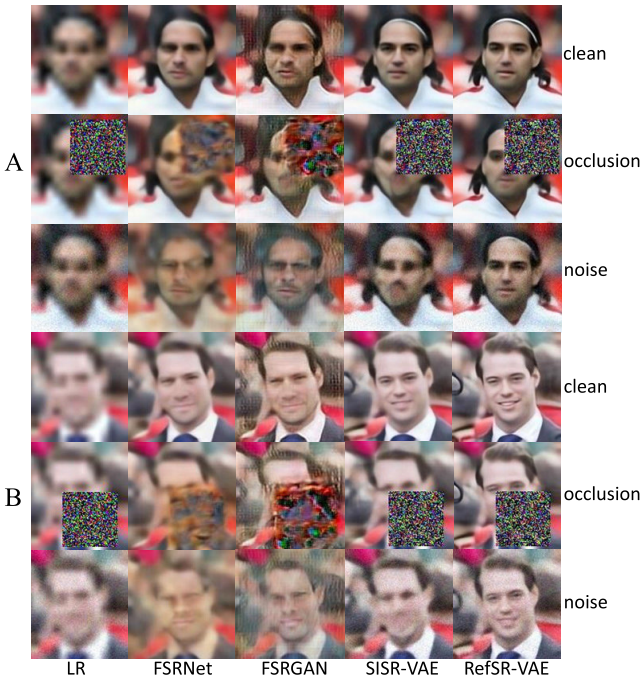

FIGURE 11. Visual comparison of denoising and occlusion.

TABLE 5. PSNR and SSIM evaluation on LFW dataset.

Eval.	PSNR(dB)	SSIM
Bicubic	22.16	0.591
SRCNN	22.45	0.591
FSRNet	20.48	0.624
FSRGAN	18.55	0.515
SISR-VAE	25.11	0.720

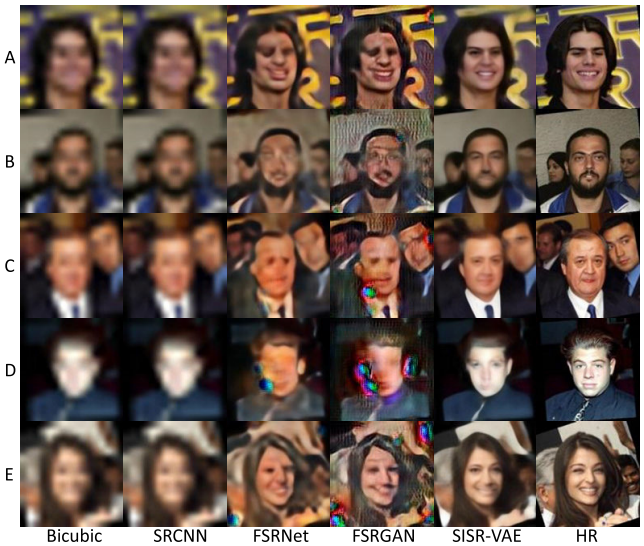

FIGURE 12. Visual comparison on LFW dataset.

In Fig. 12 we further show the visualization of different SR algorithms on the LFW dataset. Similar to the previous experiments on noisy and occluded LR images, we can observe the same problems on FSRNet and FSRGAN SR results. The colors and features on SR images are severely distorted. The reason is that LFW images are photos taken from daily life or snapshot from magazines or books. The images are filled with random noise and inference. FSRGAN and FSRNet were built on the generative adversarial network to learn the generation of photo-realistic images. The vulnerability of GAN has been discussed in some research works [49], [50]. Adversarial attack is one emerging topic of GAN and researchers study different attacks to prevent the collapse of models. The possible explanation is that it is due to insufficient model averaging and insufficient regularization so that the generated images are easily affected by random noise or some types of inference. However, as introduced in Section III, our proposed SISR-VAE makes use of the encoder to statically learn the compact latent parameters of images to get close to random Gaussian distribution. The random sampling mechanism is embedded in the training process to learn a robust decoder for image SR. Hence, SISR-VAE can resist the random noise and output good SR results.

E. ANALYSIS OF FACE IDENTITY TRANSFER

Finally, we extend the study of RefSR to face transfer, which we make use of proposed RefSR-VAE model to achieve the goal of face identity transfer from one person to another. Face generation or face attribute manipulation can be considered as domain specific of image generation. The purpose of image generation is to learn a parametric model of the training data. A good generative model should be able to ''understand'' the training data and give new samples that mix up various data attributes.

Similarly, our proposed RefSR-VAE model can also be used for style transfer. Generally, style transfer is HR-to-HR translation to generate photo-realistic images. On the other hand, our proposed RefSR-VAE not only transfers the attributes of reference image to LR images, but also perform super-resolution. We randomly selected several reference images and LR images from RefSR-Face dataset, and used the trained RefSR-VAE model to perform style transfer. Note that the RefSR-VAE is the same model as RefSR-VAE(reference) used in the last part. We have the results in Fig. 13.

In Fig. 13, we used 6 LR images and 6 corresponding reference images to conducted the experiments. Each input was a combination of one LR and one reference image so we obtained 36 different SR results. The images on the diagonal line are the SR results that used reference images of same identities. Hence, the images in red windows give the best visual quality with sharp facial features. For the results on the non-diagonal positions, they mix up the facial attributes of LR and reference images. For example, Ref_A image is a smiling woman with open mouth. All SR images are able to catch this character even the people in the LR images (LR_D and LR_F) close mouth. Though Ref_B and Ref_C are photos of women. The model can smoothly transfer female features of the reference images into the images of men (LR_D,

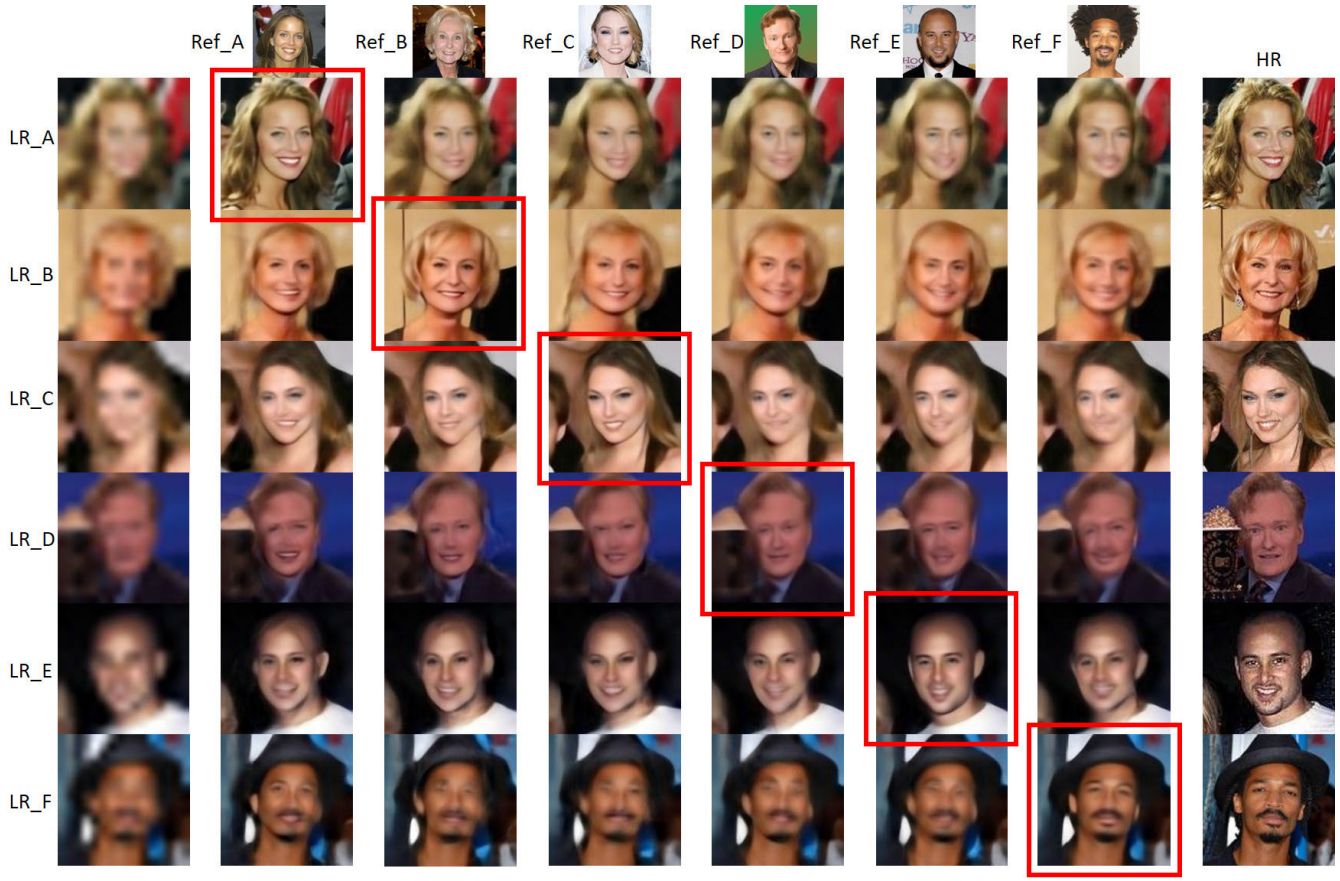

FIGURE 13. Visual comparison of facial attributes transfer.

LR_E and LR_F). On contrary, Ref_D, Ref_E and Ref_F are photos of men. The model can masculinize the photos of women (LR_A, LR_B and LR_C) to show corresponding male features of the references. Especially, Ref_F is a man with mustache, mustache-like feature can be converted to SR images.

In summary, we conducted several experiments to compare our proposed SISR-VAE and RefSR-VAE with other face SR algorithms on different dataset. From both quantitative and qualitative evaluation, our proposed approaches can achieve good SR results.

V. CONCLUSION

We have proposed a Reference based face SR via conditional Variational AutoEncoder (RefSR-VAE) and a Single Image Super-Resolution via conditional Variational AutoEncoder (SISR-VAE) for face super-resolution. Our proposed works are built on the variational autoencoder to learn the generative model of joint distribution of low-resolution and high-resolution (or reference) images for image reconstruction. The encoder is able to extract the latent parameters from the training images so that the decoder can generate new data point from learned distribution. Due to the large amount of information loss in 8× SR, we propose the Reference based face SR (RefSR) to extract useful information from the reference images to assist getting quality SR. We also come up with a novel reference based SR face dataset (RefSR-Face) to develop and test the RefSR. From a large number of experiments and the comparison with the state-of-the-art SR algorithms on different face datasets, it is found that our proposed approaches are effective and robust. Especially, our RefSR-VAE is able to provide good SR images with the highest PSNR and sharp photo-realistic quality. Furthermore, we also discover the potential use of RefSR-VAE for facial style transfer that can be used for facial expression transfer and facial image generation.

In the future work, researchers may focus on robust variational autoencoders for better face SR. There are many regularization-based methods and structured priors that can be embedded into variational autoencoders to learn constrained latent representations to generate sharp images. With the ability of generative learning of VAE models, facial image generation and completion become even a more interesting topic. This is really a fruitful direction for future research.

REFERENCES

[1] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, Mar./Apr. 2002.

[2] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Washington, DC, USA, Jun. 2004, p. 1.

[3] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sydney, NSW, Australia, Dec. 2013, pp. 1920–1927.

[4] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. IEEE Int. Conf. Asian Conf. Comput. Vis. (ACCV)*, Singapore, Apr. 2015, pp. 111–126.

[5] Z.-S. Liu, W.-C. Siu, and J.-J. Huang, "Image super-resolution via hybrid NEDI and wavelet-based scheme," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2015, p. 1.

[6] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4220–4231, Oct. 2014.

[7] 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.

[8] R. A. Farrugia and C. Guillemot, "Face hallucination using linear models of coupled sparse support," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4562–4577, Sep. 2017.

[9] J. Shi, X. Liu, Y. Zong, C. Qi, and G. Zhao, "Hallucinating face image by regularization models in high-resolution feature space," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2980–2995, Jun. 2018.

[10] J.-J. Huang and W.-C. Siu, "Learning hierarchical decision trees for single-image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 99, pp. 937–950, Dec. 2015.

[11] J.-J. Huang, T. Liu, P. L. Dragotti, and T. Stathaki, "SRHRF+: Self-example enhanced single image super-resolution using hierarchical random forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1067–1075.

[12] Z.-S. Liu, W.-C. Siu, and J.-J. Huang, "Image super-resolution via weighted random forest," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Toronto, ON, Canada, Mar. 2017, pp. 1019–1023.

[13] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, "Fast image super-resolution via Randomized Multi-split Forests," in *Proc. IEEE Int. Conf. Symp. Circuits Syst. (ISCAS)*, Baltimore, MD, USA, May 2017, pp. 1–4.

[14] L. Zhi-Song and W.-C. Siu, "Cascaded random forests for fast image super-resolution," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 2531–2535.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Sep. 2014, pp. 1–12.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, Dec. 2014, pp. 2672–2680.

[18] 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. 2016.

[19] J. Kim, J. KwonLee, and K. MuLee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1646–1654.

[20] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1637–1645.

[21] C. Ledig, L. Theis, F. Huszar, 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 Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4681–4690.

[22] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 136–144.

[23] M. Haris, G. Shakhnarovich, and N. Ukita, " Deep back-projection networks for super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 1664–1673.

[24] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, "Joint back projection and residual networks for efficient image super-resolution," in *Proc. IEEE Int. Conf. APSIPA Annu. Summit Conf. (APSIPA)*, Honolulu, HI, USA, Nov. 2018, pp. 1054–1060.

[25] Z.-S. Liu, L.-W. Wang, C.-T. Li, and W.-C. Siu, "Hierarchical back projection network for image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019.

[26] L. A. Gatys, A. S. Ecker, and M Bethge, "A neural algorithm of artistic style," Aug. 2015, *arXiv:1508.06576*. [Online]. Available: https://arxiv.org/abs/1508.06576

[27] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, and Y. Qiao, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. IEEE Eur. Conf. Comput. Workshops (ECCVW)*, Munich, Germany, Sep. 2018.

[28] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Sep. 2016, pp. 318–333.

[29] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Sep. 2016, pp. 614–630,

[30] O. Tuzel, Y. Taguchi, and J. R. Hershey, "Global-local face upsampling network," Mar. 2016, *arXiv:1603.07235*. [Online]. Available: https://arxiv.org/abs/1603.07235

[31] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3760–3768.

[32] 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)*, Murich, Germany, Sep. 2018, pp. 217–233.

[33] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Murich, Germany, Sep. 2018, pp. 183–198.

[34] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang, "Learning warped guidance for blind face restoration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Murich, Germany, Sep. 2018, pp. 272–289.

[35] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "FSRNet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2492–2501.

[36] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7982–7991.

[37] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–25.

[38] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6228–6237.

[39] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Calgary, AB, Canada, Dec. 2014, pp. 1–14.

[40] D. P. Kingma, T. Slimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Montreal, MD, Canada, Dec. 2015, pp. 3483–3491.

[41] J. Engel, M. Hoffman, and A. Roberts, "Latent constraints: Learning to generate conditionally from unconditional generative models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr. 2018, pp. 1–22.

[42] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, Xi'an, China, May 2018, pp. 67–74.

[43] V. Le, J. Brandt, Z. Lin, and T. Huang, "Interactive facial feature localization," in *Proc. IEEE Eur. Conf. Comput. Vis. (ECCV)*, Firenze, Italy, 2012, pp. 679–692.

[44] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, CN, USA, Dec. 2015, pp. 3730–3738.

[45] Y. Li, S. Liu, J. Yang, and M. Yang, "Generative Face Completion," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3911–3919.
[46] M. Mäkitalo and A. Foi, "Spatially adaptive alpha-rooting in BM3D sharpening," in *Proc. SPIE Electron. Imag., Image Process., Algorithms Syst. IX*, San Francisco, CA, USA, vol. 7870, Jan. 2011.
[47] W. Dong, G. Shi, X. Hu, and Y. Ma, "Nonlocal sparse and low-rank regularization for optical flow estimation," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4527–4538, Oct. 2014.
[48] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," in *Advances in Face Detection and Facial Image Analysis*. Cham, Switzerland: Springer, 2016, pp. 189–248.
[49] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.
[50] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr. 2018, pp. 1–17.

ZHI-SONG LIU received the M.Sc. degree in electronic engineering from The Hong Kong Polytechnic University, Hong Kong, in 2015, where he is currently pursuing the Ph.D. degree under the supervision of Prof. W.-C. Siu and Dr. Y.-L. Chan. His research interests include deep learning techniques, image and video signal processing, and image and video super-resolution.

WAN-CHI SIU (S'77–M'77–SM'90–F'12–LF'16) received the M.Phil. degree from The Chinese University of Hong Kong, in 1977, and the Ph.D. degree from Imperial College London, in 1984. He was a Chair Professor, the Founding Director of the Signal Processing Research Centre, the Head of Electronic and Information Engineering Department, and the Dean of the Engineering Faculty, The Hong Kong Polytechnic University, where he is currently an Emeritus Professor. He is also an Expert in DSP, transforms, fast algorithms, machine learning, and conventional and deep learning approaches for super-resolution imaging, 2-D and 3-D video coding, and object recognition and tracking. He has published 500 research articles (over 200 appeared in international journal articles) and edited three books. He has also nine recent patents granted. He is also a member of the IEEE Educational Activities Board, the IEEE Fourier Award for Signal Processing Committee, and some other IEEE committees. He is also a Fellow of IET and the Immediate-Past President (for the period 2019–2020) of the Asia-Pacific Signal and Information Processing Association (APSIPA). He is an outstanding scholar, with many awards, including the Best Teacher Award, the Best Faculty Researcher Award (twice), and IEEE Third Millennium Medal, in 2000. He has been a Guest Editor/Subject Editor/AE of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *Electronics Letters* and has organized, very successfully, over 20 international conferences, including the IEEE society-sponsored flagship conferences, such as the TPC Chair of the ISCAS1997 and the General Chair of ICASSP2003 and ICIP2010. He was the Vice-President, the Chair of Conference Board, and a Core Member of Board of Governors of the IEEE Signal Processing Society, from 2012 to 2014. He was an Independent Non-Executive Director of a publicly listed video surveillance company, from 2000 to 2015, and a Convenor of the First Engineering/IT Panel of the RAE (1992/1993) in Hong Kong.

YUI-LAM CHAN (S'94–A'97–M'00) received the B.Eng. (Hons.) and Ph.D. degrees from The Hong Kong Polytechnic University, Hong Kong, in 1993 and 1997, respectively.

He joined The Hong Kong Polytechnic University, in 1997, where he is currently an Associate Professor with the Department of Electronic and Information Engineering. He is also actively involved in professional activities. He has authored over 110 research articles in various international journals and conferences. His research interests include multimedia technologies, signal processing, image and video compression, video streaming, video transcoding, video conferencing, digital TV/HDTV, 3DTV/3DV, multiview video coding, machine learning for video coding, and future video coding standards, including screen content coding, light-field video coding, and 360ř omnidirectional video coding. He was the Special Sessions Co-Chair and the Publicity Co-Chair of the 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference and the Technical Program Co-Chair of the 2014 International Conference on Digital Signal Processing. He was the Secretary of the 2010 IEEE International Conference on Image Processing. He serves as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING.