Photo-Realistic Image Super-Resolution via Variational Autoencoders

Zhi-Song Liu, *Student Member, IEEE,* Wan-Chi Siu, *Life Fellow, IEEE,* and Yui-Lam Chan, *Member, IEEE*

***Abstract*—There is a great leap in objective accuracy on image super-resolution, which recently brings a new challenge on image super-resolution with larger up-scaling (e.g. 4×) using pixel based distortion for measurement. This causes over-smooth effect which cannot grasp well the perceptual similarity. The advent of generative adversarial networks makes it possible super-resolve a low-resolution image to generate photo-realistic images sharing distribution with the high-resolution images. However, generative networks suffer from problems of mode-collapse and unrealistic sample generation. We propose to perform Image Super-Resolution via Variational AutoEncoders (SR-VAE) learning according to the conditional distribution of the high-resolution images induced by the low-resolution images. Given that the Conditional Variational Autoencoders tend to generate blur images, we add the conditional sampling mechanism to narrow down the latent subspace for reconstruction. To evaluate the model generalization, we use KL loss to measure the divergence between latent vectors and standard Gaussian distribution. Eventually, in order to balance the trade-off between super-resolution distortion and perception, not only that we use pixel based loss, we also use the modified deep feature loss between SR and HR images to estimate the reconstruction. In experiments, we evaluated a large number of datasets to make comparison with other state-of-the-art super-resolution approaches. Results on both objective and subjective measurements show that our proposed SR-VAE can achieve good photo-realistic perceptual quality closer to the natural image manifold while maintain low distortion.**

***Index Terms*—Image super-resolution, variational autoencoders, distortion, divergence**

I. Introduction

IMAGE Super-Resolution (SR) is a fundamental problem in image processing. It is widely used in various image applications. With the advent of high-definition and ultra-definition images and videos, fast and accurate image super-resolution can ease the burden of broadcast and storage.

Given a LR image, the goal of image super-resolution is to resolve the following Maximum A Posterior (MAP) problem,

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

where $\hat{\mathbf{Y}}$ is the predicted super-resolution (SR) image, $\mathbf{Y}$ is the high-resolution (HR) image, $\mathbf{X}$ is the low-resolution (LR) image, $\log p(\mathbf{X}|\mathbf{Y})$ 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 model optimization. Formally, we resolve the image SR problem as follows,

$$\hat{\mathbf{Y}} = \arg\min_{\mathbf{Y}} \frac{1}{2}\|\mathbf{X} - \mathbf{C}\mathbf{Y}\|^2 + \lambda\Omega(\mathbf{Y}) \tag{2}$$

where $\mathbf{C}$ is the mapping model and $\hat{\mathbf{Y}}$ is optimized HR image given the regularization term $\Omega(\mathbf{Y})$.

The way to resolve Eq. (2) can be categorized into two major categories, i.e., generative model based and discriminative model based methods. For discriminative model based methods [1]-[14], the basic idea is to directly learn the prior parameters θ from the training data $\mathbf{Y}$ as a MAP estimation $p(\mathbf{Y}|\mathbf{X},\theta) = p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y};\theta)$. Generally, we can have the minimization optimization as

$$\min_{\theta}\|\mathbf{Y} - \hat{\mathbf{Y}}\|^r \text{ s.t.} \hat{\mathbf{Y}} = \arg\min_{\mathbf{Y}} \frac{1}{2}\|\mathbf{X} - \mathbf{C}\mathbf{Y}\|^2 + \lambda\Omega(\mathbf{Y};\theta) \tag{3}$$

where $\|*\|^r$ represents the r-th order estimation of pixel based distortion. In general, researchers assume that the residual $\epsilon = \mathbf{Y} - \hat{\mathbf{Y}}$ is independent of $\mathbf{Y}$ that follows random Gaussian distribution, and they can use r=2 to form the minimum mean squared error (MMSE) estimation. Based on the assumption that natural images are formed by a set of small overlapping image patches so that small patches in both LR and HR images form manifolds with similar local geometry in two distinct spaces, researchers can use a smaller training image set to extract more patch patterns to discover low-dimensional manifolds in high-dimensional space and embed them onto low-dimensional spaces using neighbourhood patches. For using a fixed number of neighbors for reconstruction, the SR results suffer from the blurring effect. To resolve the problem, researchers propose to improve the SR quality in two aspects: using stricter regularization for model optimization (e.g. sparsity) and using more accurate and fine grained patch matching (e.g. random forests). For the former, sparse representation based approaches [3]-[5] learn the overcomplete coupled dictionary for robust LR-HR patch mapping. For the latter, researchers make use of various classification tools [7]-[13] to classify the external and internal patches in a greedy manner, down to the maximum clusters. However, image SR quality is not completely correlated with MSE, especially for SR of large up-scaling factors.

Perceptual quality, on the other hand, focuses on divergence between SR and ground truth image distributions.

Zhi-Song Liu is with the Center Multimedia Signal Processing, Department of Electronic and Information Engineering (EIE), Hong Kong Polytechnic University, Hung Hum, Hong Kong. 100077.
E-mail: zhisong.ra.liu@ connect.polyu.hk.

Wan-Chi Siu is with the Center Multimedia Signal Processing, Department of Electronic and Information Engineering (EIE), Hong Kong Polytechnic University, Hung Hum, Hong Kong. 100077.
E-mail: enwcsiu@polyu.edu.hk.

Yui-Lam Chan is with the Center Multimedia Signal Processing, Department of Electonic and Information Engineering (EIE), Hong Kong Polytechnic University, Hung Hum, Hong Kong. 100077.
E-mail: enylchan@polyu.edu.hk.

$$\min_{\theta} d(p_{\hat{\mathbf{Y}}}, p_{\mathbf{Y}}) \text{ s.t.} \hat{\mathbf{Y}} = \arg\min_{\mathbf{Y}} \frac{1}{2}\|\mathbf{X} - \mathbf{C}\mathbf{Y}\|^2 + \lambda\Omega(\mathbf{Y};\theta) \quad (4)$$

In Eq. (4), $d(*)$ represents the divergence between distributions, e.g. Kullback-Leibler (KL) divergence, Total-Variation (TV), etc. Different from Eq. (3), using $d(*)$ measures the semantic similarity rather than pixel based errors. Based on the divergence of image distributions, there are many generative model based approaches [27, 30, 35] aiming at good perceptual quality over distortion. It is a common observation in image SR with large up-scaling factors that a SR result with high PSNR usually suffers from oversmoothing effect on visual quality. On the other hand, using generative model for SR can achieve sharp visual quality but poor PSNR performance. Due to the fact that GAN based models need to minimize the generator and maximize the discriminator iteratively, they may suffer from the lack of data diversity or unrealistic generation. This dilemma between perception and distortion can be eased by our proposed SR-VAE via minimizing the model divergence and pixel errors together. Fig. 1 shows the case of super-resolving a LR image by 4 times using the discriminative approach (EDSR [19]), generative approach (ESRGAN [35]) and our proposed Image Super-Resolution via Variational AutoEncoders (SR-VAE) for comparison in terms of PSNR and PI (one of quantitative perception measurements, with lower PI means better visual quality). We can observe that using ESRGAN, it generates a sharper SR result (containing bizarre pattern on the pedestrian and ground, and color distortion) but lower PSNR compared to EDSR (even lower than Bicubic in this case) while EDSR generates a SR image with the highest PSNR but blurry visual quality. However, we can find that our approach can achieve photo-realistic visual quality comparable to ESRGAN (e.g., clear textures around the ground and pedestrian.) and high PSNR comparable to EDSR. In Fig. 1, a light color image is used which can show easily the effect of our approach; however, full color images are mainly used in the Experiments section of the paper for comparisons and illustration.

Instead of learning the direct mapping between LR and HR images, the idea of generative model is to model the likelihood of LR images given HR images and the prior probability of the HR images, and then the Bayes rule is applied to generate reconstructed SR images. The Bayesian algorithm can naturally be used for image SR by exploiting the Gaussian process prior. Since the true distribution of HR images is unknown, researchers came up with many statistical models to approximate it, including Gaussian Mixture Model (GMM) [6, 13] and Hidden Markov Model (HMM) [9]. Recently, Generative Adversarial Network (GAN) [26] and Variational AutoEncoders (VAE) [25] have been proposed, which make use of convolutional neural networks to avoid pre-determined mixture models for the approximation and to directly train the network. This is to learn the image statistics by sampling from a semantically structured latent space. GAN based SR approaches [27, 30, 35] have achieved significant performance in perceptual quality. The adversarial network can implicitly estimate the latent parameters by expressing the input and

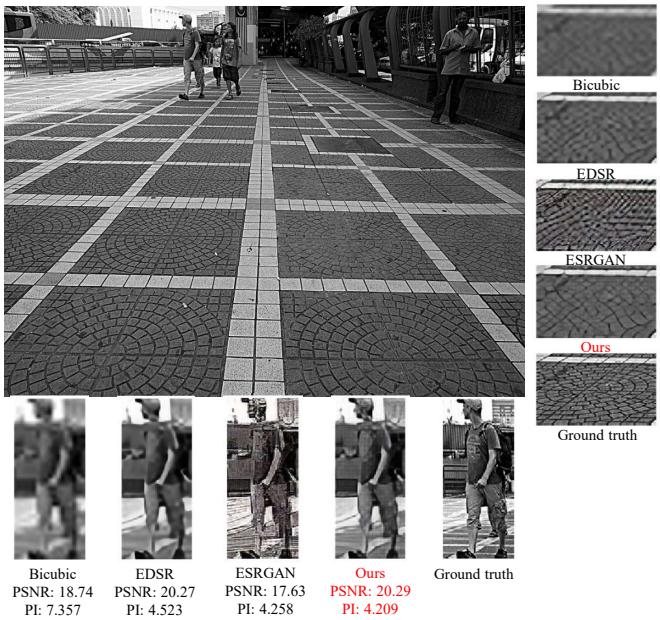

Fig. 1. Comparison among discriminative, generative and our proposed approaches on 4× image SR on *img095* from Urban100 dataset.

target as a conditional distribution on a label $c \in \{0, 1\}$. Despite the hot investigation on the GAN, there are four possible disadvantages. 1) Sensitive to hyper parameters: i.e., the training is difficult when the discriminator is trained too well and the training gradient for the generator would vanish. 2) Challenging for evaluation on the generalization measurement: i.e., GAN may overfit on some dataset hence it fails to be used in practical application. 3) Mode collapse [28]: i.e., GAN does not fully capture the diversity of the true distribution. 4) Unrealistic sample generation [31]-[32]. i.e., GAN for image SR needs to exploit the feature space based loss rather than the pixel based mean squared errors to encourage the reconstruction quality on perception over distortion. On the other hand, VAE has also attracted a lot of attentions in the past few years. It uses the autoencoder to explicitly estimate the latent parameters by maximizing a lower bound on the log-likelihood. Its mathematical formulation ensures the tractable likelihood for evaluation and explicit inference network. More details are further explained in the next section. The difficult part of VAE is the prior estimation of the latent variables for better visual quality. [45] proposes PixelVAE to hierarchically learn a better representation to improve image details. [46] proposes a novel discrete variational representation by combining vector quantization and autoencoder (Vector Quantized Variational AutoEncoder) for high-quality image generation. The results show much better performance compared to GAN based approaches. These studies indicate the huge potential of VAE for image processing.

In this paper, we propose to perform Image Super-Resolution via Variational AutoEncoders (SR-VAE) to generate photo-realistic images. Our contributions include the following points.

- To the best of our knowledge, this is the first work that

successfully makes use of the VAE network for image SR.

- **Conditional sampling mechanism**. To overcome the blurring effect in VAE sample generation, we propose a novel sampling process to sample from LR images rather than using the random Gaussian sampling. By utilizing the "reparameterization trick", we can learn conditional latent parameters for super-resolution reconstruction.
- **Back projection based SR-VAE network**. Our proposed SR-VAE network contains three subnets: SR encoder, SR decoder and VGG feature extractor. The encoder compresses the LR-HR data to learn the conditional distribution. The VGG feature extractor explores the feature similarity to reduce the divergence between HR and SR distributions. The decoder adopts the back projection based residual blocks to learn hierarchical features in an up- and down-sampling manner to reduce the residues for better reconstruction.
- **Modified VGG feature based loss estimation**. To learn the feature similarity, we adopt the same idea of GAN network to use pre-trained VGG network to extract the feature maps. We have replaced the Maxpooling opeartion in VGGNet to the Average Pooling operation to avoid pixel misalignment so that the training can be more stable. The testing results also show that using the modified VGG network can obtain smoother results without any fake feature.
- **Objective and subjective analysis**. With the success of deep learning based SR benchmarks and challenges (e.g. NTIRE [33] and PIRM [34]), many state-of-the-art approaches have been proposed to make comparison using both fully-reference measurement (e.g. Mean Squared Error) for distortion evaluation and no-reference measurement (e.g. NIQE, proposed in [29]) for perception evaluation. In our experimental work, we have used these two very different measurements for different datasets, and found that the proposed SR-VAE can achieve better perceptual quality with lower distortion.

II. Related Work

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

Perceptual measurement is an interesting topic in no-reference image quality measurement. It describes the degree of a SR image looking like a natural image with semantic similarity. One recent research work [37] investigates the relationship between perception and distortion of image restoration and finds that there is a tradeoff between them. Higher perception usually is at the expense of high distortion. Though this relationship is yet to be further proven mathematically, the authors analyzed and summarized the perception and distortion based quality measurement tools on various image SR approaches to show the tradeoff experimentally. It would be interesting to make reference to their work to study the perceptual measurement for image SR.

Generative Adversarial Networks (GAN) have inspired fruitful research works in the past few years. The idea is to train two networks (generator and discriminator) against each other to maximize the likelihood of output given some random latent vectors. The generator focuses on generating realistic samples to fool the discriminator and the discriminator learns to distinguish the fake samples from the real samples. For the generator, most of GAN based image SR approaches directly use deep neural networks, like the ResNet [15] and state-of-the-art image SR networks, such as EDSR [19], DBPN [20] and HBPN [21]. For the discriminator, it is a simple network which reduces the dimension of the input data using convolution process with larger strides to output the predicted label for 0-1 classification. A milestone of using GAN for image SR is the Super-Resolution using GAN (SRGAN) [27]. To reconstruct the texture details, the authors proposed to use the perceptual loss that applies the L2 loss over the feature maps generated from the pre-trained network (e.g. VGGNet [14]). Inspired by SRGAN, [35] proposes an Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) that uses a relativistic discriminator to distinguish real samples from the average expectation of the fake samples as a soft-decision classification to handle unrealistic sample generation. To train a GAN model, it needs to alternatively switch between the generator and discriminator to achieve the goal of minimization of generator loss and maximize the discriminator loss. There are two major problems of GAN based approaches: 1) the minimax causes gradient oscillation which increases the training difficulties, and 2) it also causes the problem of unnatural "fake" features (say the pedestrian and ground regions of using ESRGAN in Fig. 1).

Variational AutoEncoder for image generation is another choice for generative CNN model. The idea is to use the convolution neural network to explore the data distribution by solving the variational approximation to maximum-likelihood estimation. The encoder part uses the convolution process to map the input data to the latent variables for KL divergence minimization. The decoder randomly samples from the variational approximation model to output new samples. In [38], the authors came up with facial image generation via VAE network that proves the efficiency of VAE. Later on, to resolve the blurring effect on VAE generation, [39] proposes the latent constraints on the sampling procedure of the decoder to enforce the similarity of the data similarity. Furthermore, there are also some works [40]-[43] making use of conditional variational autoencoders for conditioned sample generation. To the best of our knowledge, there is no VAE based work on image SR. Inspired by VAE related work, we propose the SR-VAE via the conditional VAE network to generate photo-realistic images with sharp visual quality. The training process is very efficient and stable compared to the GAN based approaches.

III. Proposed Approach

We now formally introduce the Conditional Variational Autoencoders for image super-resolution. The whole structure as shown in Fig. 2 includes four parts: LR-HR Encoder, SR Decoder, Conditional Sampling Generator and modified VGG feature extractor. We denote a RGB LR image by

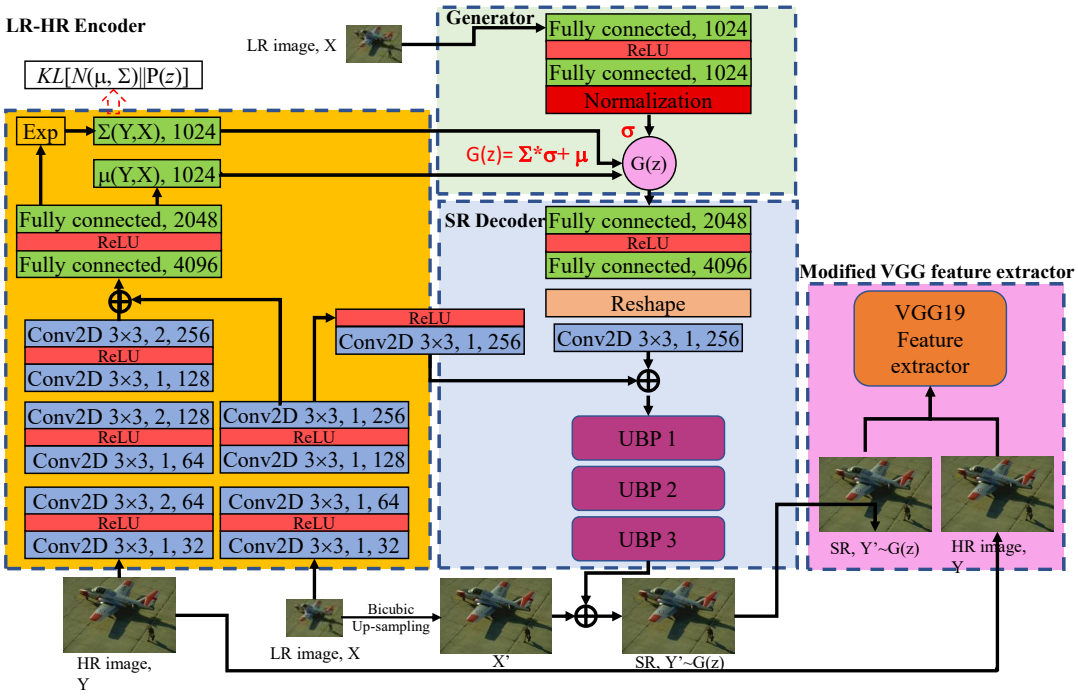

Fig. 2. Overall training structure of SR-VAE.

$\mathbf{X} \in \mathbb{R}^{m \times n \times 3}$, and its HR image by $\mathbf{Y} \in \mathbb{R}^{\alpha m \times \alpha n \times 3}$, where $(m,\ n)$ is the dimension the image and α is the up-sampling factor. To build the generative model, latent variables z in a high-dimensional space $\mathbf{Z}$ are used in our discussion to decide the hidden states of the images. Formally, for each image $\mathbf{Y}$, there is a vector of latent variables z which can randomly sample according to the probability density function (PDF) $P(z)$ and parameterized by a vector θ. The entire generative process can mathematically be described as,

$$P(\mathbf{Y}) = \int P(\mathbf{Y}|z;\theta)P(z)dz \tag{5}$$

For fixed θ parameters, we can use random latent variable z from $P(z)$ to generate new and similar image $P(\mathbf{Y})$ to the training images.

A. Variational Autoencoders (VAE)

In VAE, Gaussian distribution is chosen for the generative model distribution, i.e. $P(\mathbf{Y}|z;\theta) = \mathcal{N}(\mathbf{Y}|f(z;\theta), \sigma^2 * I)$, with mean $f(z,\theta)$ and covariance $\sigma^2 * I$. To decide the latent variables z is difficult even when we consider it as multivariate normal distribution. Instead of specifying it by hand, in some VAE related works [37]-[42], researchers used Convolutional Neural Networks (CNNs) as a dimension reduction model that can map complex training data to the latent space $\mathbf{Z}$ with right statistics. It can be regarded as maximizing the likelihood of the training data with the given latent structure in the output layer. It is different from traditional Autoencoders [43] in the sense that VAE studies the latent values for various training data instead of making feature representation for posterior estimation. In order to represent the dependency of image Y on z, $P(\mathbf{Y}|z)$, we can write it as the product of $P(z|\mathbf{Y})$ and $P(z)$ by applying Bayes rule. $P(z|\mathbf{Y})$ is the ideal representation for all images. It can be approximated as $Q(z|\mathbf{Y})$, that is learned from the available training data so we can narrow down the searching space that any z is under the prior $P(z)$. This is the key of VAE that the latent distribution can be encoded and approximated by the CNNs. In order to fit the distribution, KL divergence can be used as follows,

$$D[Q(z|\mathbf{Y})\|P(z|\mathbf{Y})] = E_{z\sim Q}[logQ(z|\mathbf{Y}) - logP(z|\mathbf{Y})] \tag{6}$$

then we can resolve Equation 6 by applying Bayes rule and reform it as,

$$\begin{aligned} logP(\mathbf{Y}) - D[Q(z|\mathbf{Y})\|P(z|\mathbf{Y})] = \\ E_{z\sim Q}[logP(\mathbf{Y}|z)] - D[Q(z|\mathbf{Y})\|P((z))] \end{aligned} \tag{7}$$

To maximize $P(\mathbf{Y})$ in the left hand side, we can optimize the right hand side by making use of the stochastic gradient descent approach. $P(\mathbf{Y}|z)$ works as a decoder to reconstruct $\mathbf{Y}$ from z and $Q(z|\,Y)$ works as an encoder to encode $\mathbf{Y}$ into z. Usually, we can define $Q(z|\,Y) = \mathcal{N}(z|\mu(\mathbf{Y};\theta), \Sigma(\mathbf{Y};\theta))$. Precisely, we can maximize the following equation by finding the evidence lower bound (ELBO) $\mathcal{L}^{ELBO}$ as,

$$\begin{aligned} \mathcal{L}^{ELBO} \triangleq \frac{1}{N}\sum_n E_{z\sim Q(z|y_n)}[logP(y_n|z)] - D[Q(z|y_n)\|P(z)] \\ \leq \frac{1}{N}\sum_n logP(y_n) \end{aligned} \tag{8}$$

where N is the batch number. In Eq. (8), we can use the "reparameterization trick" [43] to randomly sample from $Q(z|Y) = \mathcal{N}(z|\mu(\mathbf{Y};\theta), \Sigma(\mathbf{Y};\theta))$ to backprop the gradient of loss through the network. Hence, given $\mu(\mathbf{Y};\theta)$ and $\Sigma(\mathbf{Y};\theta)$ of $Q(z|Y)$, we firstly sample $\epsilon \sim \mathcal{N}(0, I)$, and then compute $z = \mu(\mathbf{Y}) + \Sigma^{\frac{1}{2}}(\mathbf{Y}) * \epsilon$. We can take the gradient of Eq. (8) as,

$$\nabla\mathcal{L} = \nabla\frac{1}{N}\{\sum_{n} E_{\epsilon\sim\mathcal{N}(0,I)}[logP(y_n|z = \mu(\mathbf{Y}) + \Sigma^{\frac{1}{2}}(\mathbf{Y}) * \epsilon)] - D[Q(z|y_n)\|P(z)]\} \tag{9}$$

VAEs have similar functions as GANs that work as a deep generative model. This can generate samples that approximate the distribution of the training data. However, different from generating new samples from random variables, image SR needs to extract the input-to-output mapping relationship. We can modify VAE as a Conditional VAE to model latent variables and data, both conditioned to some random variables.

B. Conditional Variational Autoencoders

VAE can learn the distribution of training data for new sample generation. Given a set of LR images, image SR is to solve a maximum a posterior estimation. To use VAE for image SR, we propose a Conditional Variational AutoEncoder (CVAE) to study the mapping relationship between LR and HR images. The differences between VAE and CVAE can be described in the following figure.

Fig. 3(a) shows the VAE and Fig. 3(b) shows the conditional VAE. They both have the encoder to learn the latent parameters and the decoder to output new samples. The difference is that the encoder of VAE and CVAE learns different model distribution. As shown in Figure 3(b), the encoder Q jointly learn both LR and HR images to model their conditional probability as a multivariate Gaussian distribution. The decoder reconstructs the corresponding SR images given the LR images.

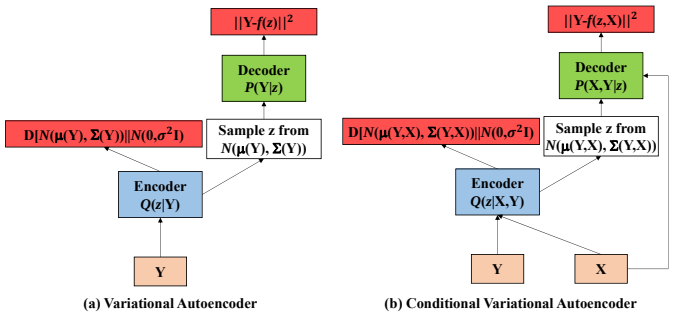

Fig. 3. Comparison between VAE network and CVAE network.

$$\frac{1}{N}\sum_{n} E_{z\sim Q(z|x_n,y_n)}[logP(x_n, y_n|z)] - D[Q(z|x_n, y_n)\|P(z)] \leq \frac{1}{N}\sum_{n} logP(y_n, x_n) \tag{10}$$

Instead of solving Eq. (8), CVAE adds the label information (Ground truth HR images) Y to model the mapping between LR and HR images. As shown in Figure 3, the CVAE model includes two loss functions: KL divergence for optimizing encoder and $l2$-norm Euclidean distance for image SR reconstruction. To sample from the learned latent variables z from $\mathcal{N}(\mu(\mathbf{X}, \mathbf{Y}), \Sigma(\mathbf{X}, \mathbf{Y}))$, we still use the "reparameterization trick" to firstly sample $\epsilon \sim \mathcal{N}(0, I)$, then compute $z = \mu(\mathbf{X}, \mathbf{Y}) + \Sigma^{\frac{1}{2}}(\mathbf{X}, \mathbf{Y}) * \epsilon$. Thus, we can rewrite Eq. (10) as following equation:

$$logP(\mathbf{X}, \mathbf{Y}|z) - D[Q(z|\mathbf{X}, \mathbf{Y})\|P(\mathbf{z})]$$
$$where\ D[Q(z|\mathbf{X}, \mathbf{Y})\|P(\mathbf{z})] = log\frac{\sigma_P^2}{\sigma_Q^2} + \frac{\sigma_Q^2 + (\mu_Q - \mu_P)^2}{2\sigma_P^2} \tag{11}$$

C. Image Super-Resolution via conditional Variational AutoEncoders (SR-VAE)

Our proposed SR-VAE network is built based on CVAE for image SR. The major modifications include three points: 1) Conditional sampling procedure, 2) Back Projection based decoder, and 3) Modified deep feature loss estimation. The complete network for 8× SR training is shown in Figure 2. The whole structure includes four parts: LR-HR Encoder, SR Decoder, Conditional Sampling Generator and modified VGG feature extractor.

LR-HR Encoder: The LR-HR Encoder takes both LR and HR images as inputs to learn the joint distribution as the output. It is made of 6 convolutional layers to gradually downscale the resolution of the feature maps. And then there are three fully connected layers to learn the latent parameters and output the k-dimensional mean vector μ(Y, X) and k-dimensional variance vector Σ(Y, X). Based on Eq. (11), the encoder needs to minimize the divergence between the target prior $P(z)$ and estimated prior $Q(z|\mathbf{X}, \mathbf{Y})$ learned from the training data. We assume prior as the normal distribution $\mathcal{N}(0, \sigma^2 * I)$ and can have the loss of the model divergence as L_D to force the training data to approximate the normal distribution.

SR Decoder: The SR Decoder is modified from our previous enhanced back projection based residual blocks [21]. Different from [21], the proposed SR Decoder is a part of the proposed SR-VAE that learns the hidden latent vector for image generation. It takes the latent vector as the input while the network in [21] takes the LR image as the input. In order to decode the latent vector, the proposed SR Decoder has two extra fully connected layers (the green boxes in Fig. 2) to project the latent vector to the spatial domain. Furthermore, there is an addition layer (the yellow box in Fig. 2) that

works as a conditional image generation to ensure that the SR image is close to the LR image. Besides the differences, as in [21], the SR Decoder also makes use of the enhanced back projection based residual blocks. We refer it to as Up-sampling Back Projection block (UBP) (as shown in Fig. 4(b)). The basic idea is to embed the back projection into the residual learning, hence it can improve the SR performance. Note that for large scale image SR, the convolution and deconvolution layers in UBP block would need large kernel sizes and strides to perform the up- and down-sampling operations. For instance, the 8× image SR needs 12×12 kernel with stride 8. Kernels with large sizes can be time consuming and the model becomes suboptimal. As discussed in VGG [14], a combination of a few convolution operations with a small kernel size can cover the same receptive filed as the one convolution with a large kernel size. We then come up with the SR Decoder by stacking multiple enhanced UBP for gradual up-sampling. Details of 8× image SR are shown in Fig. 4.

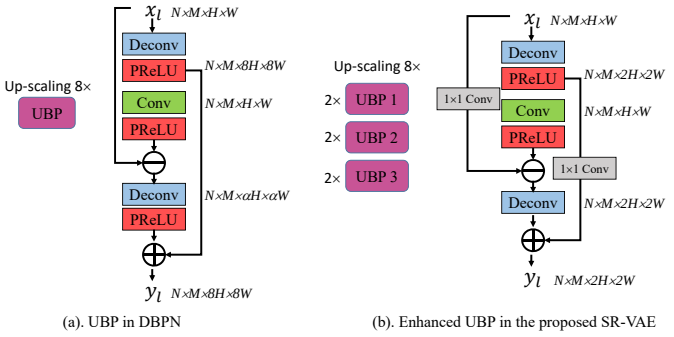

Fig. 4. Comparison of different UBP blocks. (a) the UBP block in the DBPN network. (b) the proposed enhanced UBP block in SR-VAE network. N is the batch size. M is the channel size, H and W give the size of the feature map

For $\alpha\times$ image SR, the SR Decoder includes $log_2\alpha$ UBP blocks to up-sample the feature maps. By using smaller kernels and strides, we can use more UBP blocks to build a deeper network for nonlinear mapping. Besides using smaller filers, there are a few improvements in our enhanced UBP blocks. We add two 1×1 convolutional layers (gray blocks in Fig. 4(b)) as a weighted addition for the residual update. We also remove one activation layer after the deconvolutional layer because it is redundant to add two nonlinear activated outputs for training. One nonlinear activated output with another linear convolution output can already achieve the nonlinear mapping. For comparison, we also show the back projection model in DBPN [20] (Fig. 4a). The differences can be summarized in threefold. First, our proposed enhanced UBP block uses a smaller kernel size for convolution and deconvolution. Second, the 1×1 convolution layer is added as the shortcuts to fine tune the residues for update. Third, for large scale image SR, we cascade multiple enhanced UBP blocks to gradually up-sample LR images to obtain better SR performance.

Conditional Sampling Generator: In CVAE, the "reparameterization trick" embeds the randomness of latent variables sampling in the training process to ensure the general ability. However, there is a tradeoff of sharp bizarre and blurring realistic reconstruction. Random sampling cannot manifest the key attributes of the latent variables included for reconstruction. In order to guide VAE to generate sharp reconstruction, we can build a model to learn conditional sampling to map the random set of latent variables to a subspace of the latent space. [40] introduces the constrains of sampling by adding the GAN network to transform the random variables to the conditional latent space. [39], on the other hand, uses the "Actor/Critic" training scheme to supervise the random sampling to generate realistic samples. However, for both approaches, the conditional sampling stage requires an alternative training so that the encoder and decoder compete with each other to encourage the former to learn constrained latent space and the latter to generate high-quality results.

To avoid complicated training process, we propose a Conditional Sampling Generator that adds a constraint on the latent space to enforce the latent samples to be diverse as well as realistic. We come up with a supervised sampling scheme that can train the generator to balance the trade-off between reconstruction and sample quality. The Conditional Sampling diagram is shown in Fig. 5.

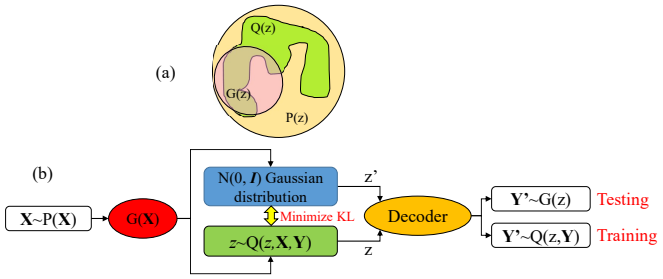

Fig. 5. Diagram of conditional sampling. (a). G(z) stands for constrained latent z. P(z) is the posterior. Q(z) is the "encoder" distribution that approximates the posterior. (b). to train the generator G, we use it to sample from "encoder" conditional distribution and pass it to the "decoder" for estimation.

Instead of randomly sampling from a normal distribution, we use generator G (details about G is shown in Fig. 2) to learn the subspace of latent variables for sampling. For generator G, it includes two fully connected layers to learn the sampling distribution. We use normalization to center the LR image pixels with mean 0 and variance 1 and then multiply it with latent variables to work the same way as the "reparameterization trick". From the perspective of image signal processing, the LR image is the low frequency components sampled from the HR image. Any data sampled by the normalized LR image can be centered to the desired frequency domain. During the training, the input LR image goes through the generator G to learn the normalized sampling distribution sampling from the "encoder" distribution (learned from ground truth images) and the decoder Y'∼Q(z,Y) learned from the training data. After training, the KL divergence between "encoder" distribution and normal distribution is minimized so that we can use G to directly sample from the normal distribution and use the decoder to output Y'∼G(z) as the prediction.

Modified VGG feature extractor: Finally, similar to GAN based SR approaches, to encourage SR-VAE generating more photo-realistic images with sharper details, we adopt the VGGNet to learn the feature differences between HR and SR

images. It is commonly used in perceptual image SR [27, 30, 35] because the feature maps at each layer of VGGnet show patterns similar to human visual perception. The details about VGG feature extractor is shown in Fig. 2. Both SR and HR images are input in the VGG network to extract their corresponding feature maps. Similar to GAN based image SR algorithms, we use the pre-trained VGG19 [14] network and fix the parameters to extract the both SR and HR feature maps $\phi_{54}(\mathbf{Y}')$ and $\phi_{54}(\mathbf{Y})$, where 54 indicates feature obtained by the 4^{th} convolution layer before the 5^{th} "Maxpooling" layer. The key difference between our modified VGG feature extractor and others is that we changed the "Maxpooling" layer in VGGNet to "Average pooling" to avoid the feature misalignment. The "Maxpooling" layer is actually an inconsistent down-sampling operation because the most activated feature points are different for HR and SR images so that during the backpropagation, the "Maxpooled" features will be placed back to different positions. With more layers of operations, the "noise" pattern can cause fluctuating change of losses in the training process. This misalignment of position can cause fake and inconsistent features on the SR results. In order to avoid the problem, we can simply use "Average pooling" to extract the average feature pixels to represent the feature responses of the sampling region. This can obtain smooth feature maps and also have steady loss minimization

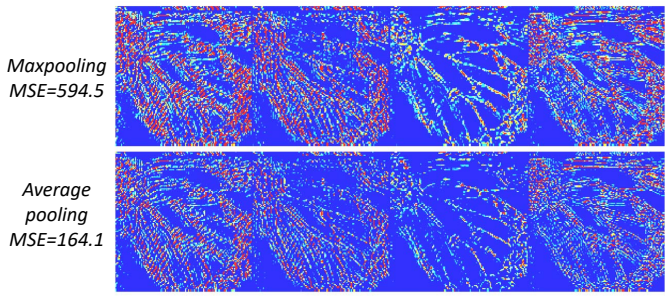

Fig. 6. Visual comparison between "Maxpooling" and "Avepooling" on image *butterfly*. The pixel value is enlarged by 10 times for better comparison.

In Fig. 6, we show the residual heatmaps generated by "Maxpooling" and "Average pooling". Blue means low activation values and red means high activation values. We take 8 feature maps from "conv12" of VGG19 network and use "Maxpooling" and "Average pooling" to down-scale the HR and SR feature maps. We can then visualize the residual features between HR and SR by heatmaps. We can see that the residual maps generated by "Maxpooling" have some ambiguous and noisy patterns around the edges, while using "Average pooling" we can obtain a clearer and smoother pattern. From the calculated MSE results, we also can see that using "Average pooling" can obtain smaller MSE compared with "Maxpooling". Both residual maps and MSE results suggest that using "Average pooling" can extract smoother feature maps with lower feature loss which can ease the training difficulty.

Finally, to balance between low distortion and high perceptual quality, the total loss is formed from a combination of the l_1-norm pixel based errors and VGG feature based errors as follow,

$$L_{SR} = \frac{1}{N}\sum_{n}\left(\left\|\mathbf{Y}-\mathbf{Y}'\right\|_1 + \lambda\left\|\phi_{54}(\mathbf{Y})-\phi_{54}(\mathbf{Y}')\right\|_1\right) \quad (12)$$

where λ stands for the balancing parameter for the VGG feature loss and L_{SR} is the loss of SR reconstruction. The first term indicates the average reconstruction distortion and the second term indicates the average perception loss. l_1-norm loss is commonly used in image SR for loss estimation. It has been widely proven useful in many works [19]-[21]. For image SR, it measures the average absolute differences between the ground truth images and the predicted SR images. It can be regarded as a first-order of the Mean Squared Errors (MSE). Experimental results show that using l_1-norm loss can achieve better performance.

Training and Testing the SR-VAE model: As shown in Eq. (13), the target is to minimize the total loss *L*, including the reconstruction loss L_{SR} and model divergence L_{KL}. For reconstruction loss, we have to balance the SR results between distortion and perception by using l_1-norm pixel based MSE and VGG feature based loss. As explained in Eq. (12), balancing parameter λ controls the importance of perception loss. Besides these two losses, we also include the KL divergence (as "KL" output of the LR-HR Encoder are shown in Fig. 2) between conditional latent and normal distribution introduced in Eq. (11). The total loss is as follows.

$$\begin{aligned} L &= L_{SR} + L_{KL} \\ where\ L_{KL} &= log\frac{\sigma_P^2}{\sigma_Q^2} + \frac{\sigma_Q^2 + (\mu_Q - \mu_P)^2}{2\sigma_P^2} \end{aligned} \quad (13)$$

where L is the total loss, including L_{SR} and L_{KL}. Unlike the alternative training process used in GAN based approaches where the parameter updating is updated in a minmax manner between generator and discriminator to reach a Nash equilibrium. The model parameters oscillate so that using gradient descent may not converge. GAN based approaches require patient hyperparameter tuning and some training tricks to balance the generator and discriminator. On the other hand, our SR-VAE training process is end-to-end and all parameters are updated at once. The KL loss and SR reconstruction loss are both optimized in a minimization manner so that they do not conflict with each other. Eventually, we have the training loss as shown in Fig. 7,

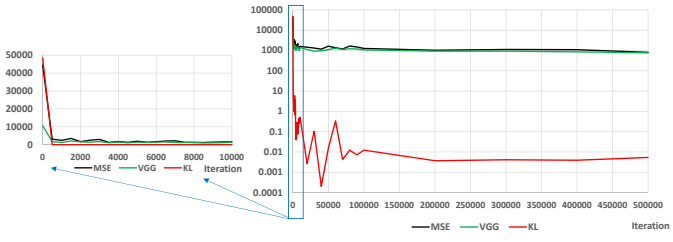

Fig. 7. Training losses of the proposed SR-VAE.

In Fig. 7, we show the relationship between the iteration time and training losses (including l_1-norm MSE loss, VGG feature loss and KL loss). Due to the large scale differences of L_{SR} and L_{KL}, we have to plot the loss in logarithmic scale. For better comparison, in Fig. 7 we have also enlarged the beginning part from 0 to 10000 iterations. We can find that all three losses are getting smaller when the number of training iterations increases.

At test time, when we want to super-resolve a LR image, we simply input the LR image into the decoder. That is, we remove the SR encoder and VGG feature extractor. The testing model is shown in the following figure.

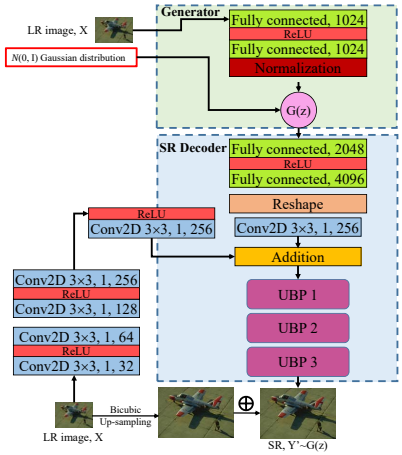

Fig. 8. Testing model of the proposed SR-VAE.

IV. Experiments

A. Implementation details

Datasets We conducted extensive experiments on various testing datasets: Set5 [51], Set14 [52], BSD100[50], Urban100 [53] and Manga109 [54], which contain images with different resolutions and contents. Generally, Set5, Set14 and BSD100 consist of images with small and medium resolutions, while Urban100 and Manga109 contain images with larger resolutions. The training images come from two commonly used datasets: DIV2K [33] and Flickr2K [19], both of which contain images with pixel resolution no smaller than 1000×1000px. During the training, the LR images were obtained by down-sampling the HR images using *Bicubic* in MATLAB, and then we used *Bicubic* again as an initial up-sampling operation to up-sample the LR images to the same dimension as the HR images.

Training Setting Based on the up-sampling factor α, we first generated LR images by using *Bicubic* and then we cropped the LR and HR images into patches of dimension 128×128 to form the LR-HR patch pairs. Similar to most SR approaches, we enlarged the number of training data by image augmentation, including flipping and rotation. Eventually, we were able to generate around 1,000,000 training LR-HR patch pairs.

Experiments were conducted on $4\times$ and $8\times$ image SR. For $4\times$ and $8\times$ SR, we used 2 UBP block and 3 UBP blocks, respectively. For SR encoder, we used ReLU activation function and Parametric ReLU (PReLU) for SR decoder. For UBP blocks, we used 6×6 filters with stride 2 for convolution and deconvolution and 3×3 filters with stride 1 for weighting operation. Pre-trained VGG19 is provided by [14]. All the weights were initialized based on [54]. We trained our model with learning rate initialized to 0.0001 for all layers and with 10 times slower after 500,000 iterations, for a total of 1,000,000 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 one NVIDIA GTX 1080 Ti GPU.

PSNR [45], SSIM[46] and Perception Index (PI) [34] were used to evaluate different settings and SR algorithms. PSNR and SSIM are very standard distortion based SR evaluation that describe the pixel based loss. On the other hand, PI has been proposed recently for perceptual quality estimation. The investigation result in [34] shows that there is a high correlation between human opinion scores and PI scores on top-10 image SR methods. This measurement approach combines the no-reference image quality measures of Ma et al. [36] and NIQE [29] as

$$PI = \frac{1}{2}\left((10 - Ma) + NIQE\right) \tag{14}$$

Note that the lower PI we get, the better perceptual quality we can obtain. As used by existing algorithms, we converted RGB images to YUV images and only used Y-channel for estimation. For SR by factor α, we excluded α pixels at the boundaries to avoid the boundary effect.

B. Analysis of Network structure

From the introduction of our proposed methods, we have indicated that there are a number of parameters needed to have further experimental analysis. We have conducted a few experiments trying to find the optimal design.

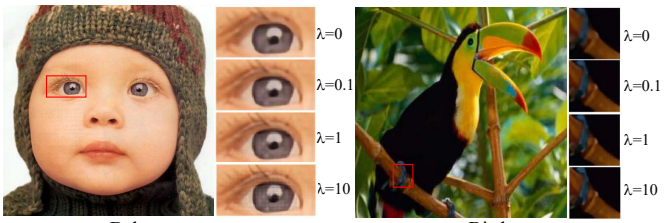

Fig. 9. Visual comparison among different λ for $4\times$ SR on Set5.

1) Effect of VGG feature loss: In the proposed SR-VAE, the effect of VGG feature loss is to control the perceptual quality. As discussed in the Introduction, distortion and Perception are at odd with each other so that we cannot guarantee a SR result with both good visual quality and low distortion. Hence, the value of the balance parameter λ in Eq. (12) is important. We

tested different values of λ and obtained the results in TABLE I.

TABLE I
COMPARISON OF USING DIFFERENT VALUES OF λ FOR 4× SR ON SET5 AND BSD100. RED INDICATES THE BEST RESULTS.

λ	Scale	Set5			BSD100		
		PSNR	SSIM	PI	PSNR	SSIM	PI
0	4	32.21	0.89	6.223	27.61	0.73	5.623
0.1	4	32.13	0.89	6.087	27.52	0.73	5.571
1	4	32.02	0.89	6.049	27.47	0.73	5.546
10	4	31.46	0.88	4.836	27.21	0.72	4.249

In TABLE I, We can find that a smaller λ can give a higher PSNR and SSIM which is reasonable because we forced the model on minimizing pixel based loss rather than VGG feature loss. However, using larger λ can generate SR images with better perceptual quality. In Fig. 9, we show the visual comparison among different λ values.

From Fig. 9, we can see that images *Baby* and *Bird* using the model with $\lambda = 10$ give sharper features compared with other setting. Note that $\lambda = 0$ means that training the SR-VAE as a MSE based model without using VGG feature loss. From TABLE I, it can give the highest PSNR and SSIM, but with rather blurry features (eyelashes in *Baby* and claws in *Bird*).

2) Effect of Conditional Sampling Generator: For our proposed SR-VAE, we used the idea of conditional VAE to super-resolve LR images. We propose to use the Conditional Sampling Generator to replace the normal distribution $\mathcal{N}(0, I)$ for conditional sampling. In TABLE II, A indicates the standard conditional VAE using normal distribution for sampling and B is the proposed SR-VAE using Conditional Sampling Generator. Besides this, in Section 3, we have introduced that the tractability of the model relies on the assumption that $Q(z|\mathbf{X}, \mathbf{Y})$ can be modeled as a Gaussian with mean $\mu(\mathbf{X}, \mathbf{Y})$ and variance $\Sigma(\mathbf{X}, \mathbf{Y})$. $P(\mathbf{X}, \mathbf{Y})$ can converge to the true distribution by looking for the lower bound of $D\left[Q(z|\mathbf{X}, \mathbf{Y})\|P(\mathbf{z})\right]$. Decreasing variance $\Sigma(\mathbf{X}, \mathbf{Y})$ of the Gaussian model can maximize the ELBO and hence increase the data fidelity. However, if the variance goes too small, it can also force the model to concentrate around the training data to generate sharp but bizarre reconstruction. In order to explore the optimal choice of the Gaussian variance, we also tested several values for comparison.

TABLE II
COMPARISON OF USING DIFFERENT VALUES OF $\Sigma(\mathbf{X}, \mathbf{Y})$ AND CONDITIONAL SAMPLING GENERATOR FOR 4× SR ON SET5 AND BSD100. RED INDICATES THE BEST RESULTS.

Model	σ	Conditional Generator	Set5			BSD100		
			PSNR	SSIM	PI	PSNR	SSIM	PI
A	0.1		31.33	0.87	5.001	27.18	0.72	4.478
	1		31.46	0.88	5.021	27.20	0.72	4.504
B	0.1	√	31.46	0.88	4.836	27.21	0.72	4.249
	1	√	31.58	0.88	4.978	27.32	0.72	4.451

In TABLE II, we can find that using the proposed Conditional Sampling Generator over normal distribution for sampling can improve the PSNR by about 0.13 dB and also decrease the PI by about 0.2. Using the same Conditional Sampling Generator, a smaller Gaussian variance ($\sigma = 0.1$) for computing the ELBO can achieve smaller PI. Note that when we were computing the ELBO, we assumed that the mean of $Q(z|\mathbf{X}, \mathbf{Y})$ and the Gaussian mean are equal to 0 for simplicity. Meanwhile, we can find that using a smaller σ can also reduce PI by about 0.04. However, computing the PSNR and PI is not enough to show the advantage of choosing Conditional Sampling Generator for perceptual quality. In Fig. 10, we show some cases of SR results with and without using Conditional Sampling Generator for comparison. We can find that using Conditional Sampling Generator can achieve more realistic and smooth features. On the other hand, the standard conditional VAE generates unrealistic features.

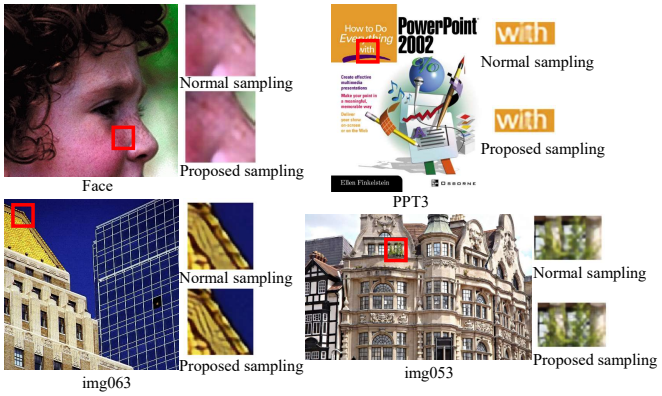

Fig. 10. Visual comparison between different SR-VAE model using with or without Conditional Sampling Generator for 4× SR.

From Fig. 10, on *Face* image, using the normal sampling to train SR-VAE model cannot capture the freckle around the nose while using proposed Conditional Sampling Generator to train SR-VAE can obtain a clearer result. On image *PPT3*, the same result can also be observed that using the proposed Conditional Sampling Generator can enhance the letter "i". For *img063* and *img053*, using the proposed Conditional Sampling Generator can recover the roof structure and leaf texture, respectively.

3) Effect of VGG feature extractor: In order to generate SR image with low distortion as well as good perceptual quality, we have added the VGG feature based loss for training. Our modified VGG feature extractor uses "Average pooling" to replace "Maxpooling" to avoid the inconsistent feature generation. This modification focuses on improving the perceptual quality that does not help to minimize image distortion. To make a comparison, we used *VGG-A* to represent the modified VGG feature extractor and *VGG-M* to represent the original VGG feature extractor. Instead of measuring PSNR or SSIM for comparison, let us directly show some SR images to visualize the effect of using *VGG-A*.

In Fig. 11, we can see that *VGG-A* can generate smoother results while *VGG-M* generates rather inconsistent features. For example, for image *8023*, the texture around the wing is

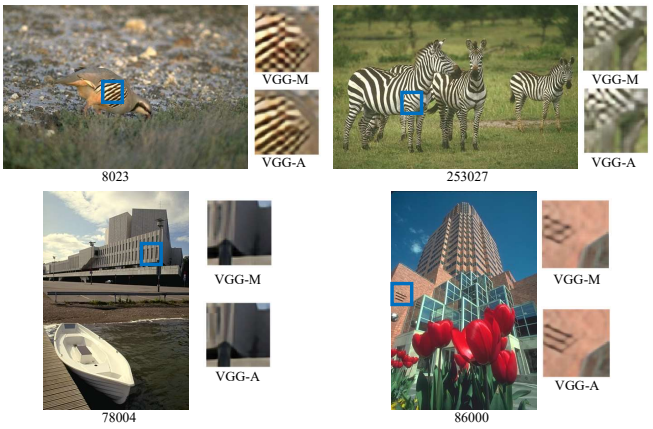

Fig. 11. Visual comparison between different SR-VAE model using "Average pooling" or "Maxpooling" for 4× SR on BSD100.

correctly reconstructed using *VGG-A*; for image *86000*, the air vent on the wall can be well recovered by using *VGG-A*.

4) Effect of UBP and DBP: In the decoder part of the proposed SR-VAE, we propose to use back projection based residual block to replace the conventional residual block to update the low- and high-resolution feature maps. The modification is that we embed the back projection mechanism into the residual block. The enhanced Up-sampling Back Projection (UBP) block follows the forward back projection process to minimize the high-resolution features while the enhanced Down-Sampling Back Projection (DBP) block conducts the inverse process of the back projection process to minimize the low-resolution features. In order to demonstrate the effectiveness of the proposed UBP and DBP blocks, we design same number of convolution layers to ensure using the same depth of feature extraction.

TABLE III
COMPARISON BETWEEN RESIDUAL BLOCK AND PROPOSED BACK PROJECTION BASED BLOCK FOR 4× AND 8× SR ON SET5 AND BSD100. RED INDICATES THE BEST RESULTS.

Scale	Conv block	Set5			BSD100		
		PSNR	SSIM	PI	PSNR	SSIM	PI
4×	ResB	31.87	0.887	6.112	27.24	0.724	5.877
	BPB	32.02	0.892	6.049	27.47	0.732	5.546
8×	ResB	27.19	0.784	7.841	24.67	0.587	7.311
	BPB	27.42	0.785	7.775	24.93	0.600	7.165

In TABLE III, we label the decoder using conventional Residual blocks as ResB and the one using UBP and DBP as BPB for clarity. We tested the SR results on 4× and 8× SR on Set5 and BSD100. We can find that using proposed UBP and DBP blocks for SR outperforms conventional Residual blocks in PSNR, SSIM and PI scores.

5) Comparison with the state-of-the-art SR algorithms: To confirm the ability of the proposed network, we performed several further experiments and analysis. Let us compare our proposed SR-VAE with nine state-of-the-art SR algorithms. Among them, there are six distortion based algorithms: A+ [5], SRCNN [16], LapSRN [18], EDSR [19], DBPN [20] and HBPN [21] and three GAN based algorithms: SRResNet [27], EnhanceNet [30] and ESRGAN [35]. For realization, A+, SRCNN, LapSRN and EDSR were reimplemented and provided by the authors of [18]. The results of DBPN, HBPN, EnhanceNet and ESRGAN are provided by respective authors. We reimplemented their released codes and generated the SR images. SRResNet was reimplemented by the authors of [35] so we used their released code to generate the SR images. Note that recent GAN based algorithms only have 4× SR results for us to make comparison. That is, there have not been 8× SR results available for us to make comparison. For the sake of fairness, we modified the official 4× SR codes of SRGAN and ESRGAN by changing the up-sampling layers and directly retrained the networks for 8× image SR without extra operations. We mark them with "*" to indicate that the modification was not done by the codes exactly provided by the original authors. For distortion based algorithms, we compare them for scale factors 4× and 8×. All the SR images can be found in https://github.com/Holmes-Alan.The evaluation was conducted based on three objective indicators: distortion, perception and running time.

In TABLE IV, we give quantitative evaluation of our approaches and the state-of-the-art SR algorithms on average PSNR, SSIM and PI for scale factors 4× and 8×. For our proposed SR-VAE, we give two results: SR-VAE-D and SR-VAE-P. SR-VAE-D represents the model trained on using balance parameter $\lambda = 1$ to give equal importance to the MSE loss and VGG feature loss. We used SR-VAE-D to generate SR images with high data fidelity, sacrificing perceptual quality. On the other hand, we also have SR-VAE-P results which were trained on using balance parameter $\lambda = 10$, to give high importance to VGG feature loss over the MSE loss. We used SR-VAE-P to generate SR images with better visual quality, sacrificing data fidelity.

In TABLE IV, we use PSNR and SSIM to evaluate the distortion and PI score to evaluate the perception. The higher PSNR and SSIM represent lower distortion and the smaller PI score represents higher perceptual quality. We can find that our proposed SR-VAE-D and SR-VAE-P can outperform GAN based approaches (EnhanceNet, SRGAN, ESRGAN) in terms of PSNR and SSIM by about 1 dB and 0.1, respectively. They can also achieve lower (about 0.5) PI scores as compared with CNN based approaches (DBPN, HBPN). However, using PI to measure the visual quality is still not objective enough for visual comparison. Hence, to make comparison of the visual quality, we include Fig. 12, 13 and 14 to compare the results to see the true visual appearance of different SR algorithms. Combining TABLE IV and Figs. 12, 13 and 14, we can have two observations:

- Comparable Perceptual quality. The PI scores in TABLE IV are an approximate evaluation of perception. The lower value we have, the better is the visual quality of the SR images. It can be found that our proposed SR-VAE-P approach (focusing more on visual quality) can achieve similar PI score to the state-of-the-art perceptual image SR approach (ESRGAN). Meanwhile, the results on Fig. 12 show that our proposed SR-VAE-P can actually generate better visual quality while other GAN based SR ap-

TABLE IV
QUANTITATIVE EVALUATION OF STATE-OF-THE-ART SR APPROACHES, INCLUDING PSNR, SSIM AND PI FOR SCALE 4× AND 8×. RED INDICATES THE BEST AND BLUE INDICATES THE SECOND BEST RESULTS.

Algorithm	Scale	Set5			Set14			BSD100			Urban100			Manga109		
		PSNR	SSIM	PI	PSNR	SSIM	PI	PSNR	SSIM	PI	PSNR	SSIM	PI	PSNR	SSIM	PI
Bicubic	4×	28.42	0.810	7.370	26.100	0.704	7.027	25.96	0.669	6.999	23.64	0.659	6.944	25.15	0.789	6.764
A+		30.300	0.859	7.010	27.43	0.752	6.788	26.82	0.710	6.457	24.34	0.720	6.115	27.02	0.850	6.001
SRCNN		30.49	0.862	7.120	27.61	0.754	6.891	26.91	00.712	6.556	24.53	0.724	6.289	27.66	0.858	6.051
LapSRN		31.54	0.885	6.497	28.19	0.772	6.210	27.32	0.728	5.808	25.21	0.756	5.596	29.09	0.890	5.312
EDSR		32.46	0.897	5.906	28.800	0.788	5.51	27.71	0.742	5.559	26.64	0.803	5.338	31.02	0.915	5.124
DBPN		32.47	0.898	6.156	28.82	0.786	5.588	27.72	0.740	5.355	26.60	0.795	5.268	31.13	0.914	5.122
HBPN		32.73	0.901	6.143	28.99	0.792	5.568	27.85	0.746	5.505	27.03	0.815	5.134	31.66	{0.920}	4.907
EnhanceNet		28.57	0.810	2.926	25.77	0.677	3.018	24.93	0.626	2.908	23.54	0.692	3.472	26.70	0.825	3.287
SRGAN		29.40	0.847	3.355	26.02	0.740	2.882	25.16	0.669	2.351	24.41	0.732	3.484	28.09	0.861	3.303
ESRGAN		30.47	0.851	3.755	26.28	0.698	2.926	25.32	0.651	2.479	24.36	0.733	3.771	28.44	0.860	3.456
SR-VAE-D (Ours)		32.02	0.892	6.049	28.30	0.776	5.229	27.47	0.732	5.546	26.28	0.792	5.183	30.87	0.912	4.747
SR-VAE-P (Ours)		31.46	0.882	4.836	27.91	0.762	4.351	27.21	0.723	4.249	26.33	0.793	4.481	30.17	0.902	4.184
Bicubic	8×	24.39	0.657	9.932	23.19	0.568	9.488	23.67	0.547	9.552	21.24	0.516	8.832	21.68	0.647	8.705
A+		25.52	0.692	9.454	23.98	0.597	8.884	24.20	0.568	8.995	21.37	0.545	8.256	22.39	0.680	8.213
SRCNN		25.33	0.689	9.956	23.85	0.593	9.056	24.13	0.565	9.231	21.29	0.543	8.412	22.37	0.682	8.402
LapSRN		26.15	0.738	9.936	24.35	0.620	7.956	24.54	0.586	8.036	21.81	0.582	7.543	23.29	0.735	7.293
EDSR		26.96	0.775	7.999	24.94	0.640	7.325	24.80	0.596	7.620	22.47	0.620	6.526	24.58	0.778	5.985
DBPN		27.21	0.784	8.285	25.13	0.648	7.419	24.88	0.601	7.660	22.69	0.622	6.618	24.96	0.799	6.050
HBPN		27.17	0.785	8.112	24.96	0.648	7.369	24.93	0.602	7.628	23.04	0.647	6.445	25.24	0.802	5.883
SRGAN*		26.30	0.755	8.000	23.98	0.0.600	7.334	24.60	0.585	7.410	22.45	0.626	6.410	24.67	0.786	5.749
ESRGAN*		26.32	0.755	8.011	24.08	0.601	7.333	24.63	0.585	7.342	22.55	0.627	6.301	24.74	0.787	5.758
SR-VAE-D (Ours)		27.42	0.785	7.775	25.03	0.638	6.851	24.93	0.600	7.165	23.19	0.651	6.125	25.37	0.800	5.599
SR-VAE-P (Ours)		26.86	0.770	7.368	24.64	0.626	6.141	24.65	0.591	6.254	22.75	0.633	5.695	24.92	0.790	5.399

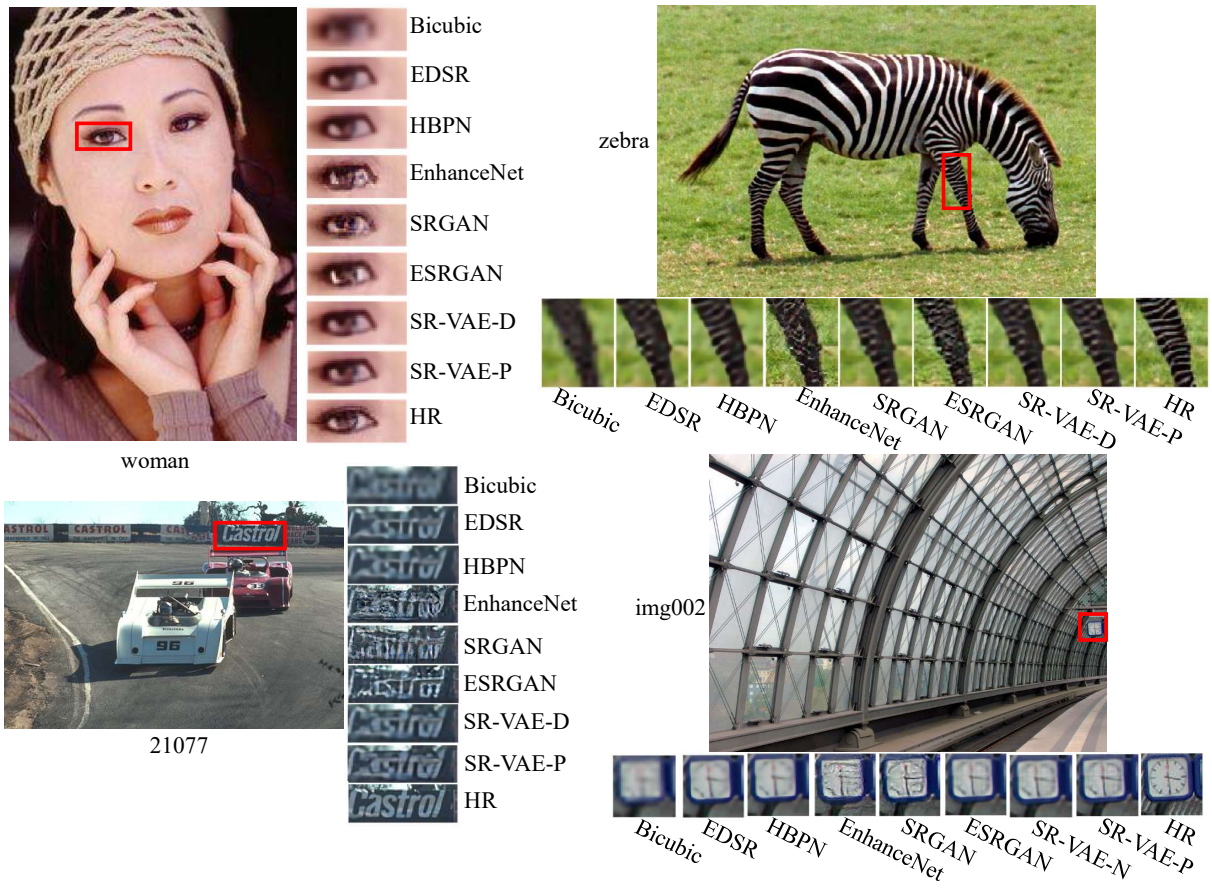

Fig. 12. Visual comparison among different SR algorithms for 4× SR.

proaches (EnhanceNet, ESRGAN and SRGAN) amplify the noise around the edges. Furthermore, with similar visual quality, our proposed SR-VAE-P can achieve higher PSNR (the higher value we obtain, the lower distortion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.3003832, IEEE Transactions on Circuits and Systems for Video Technology

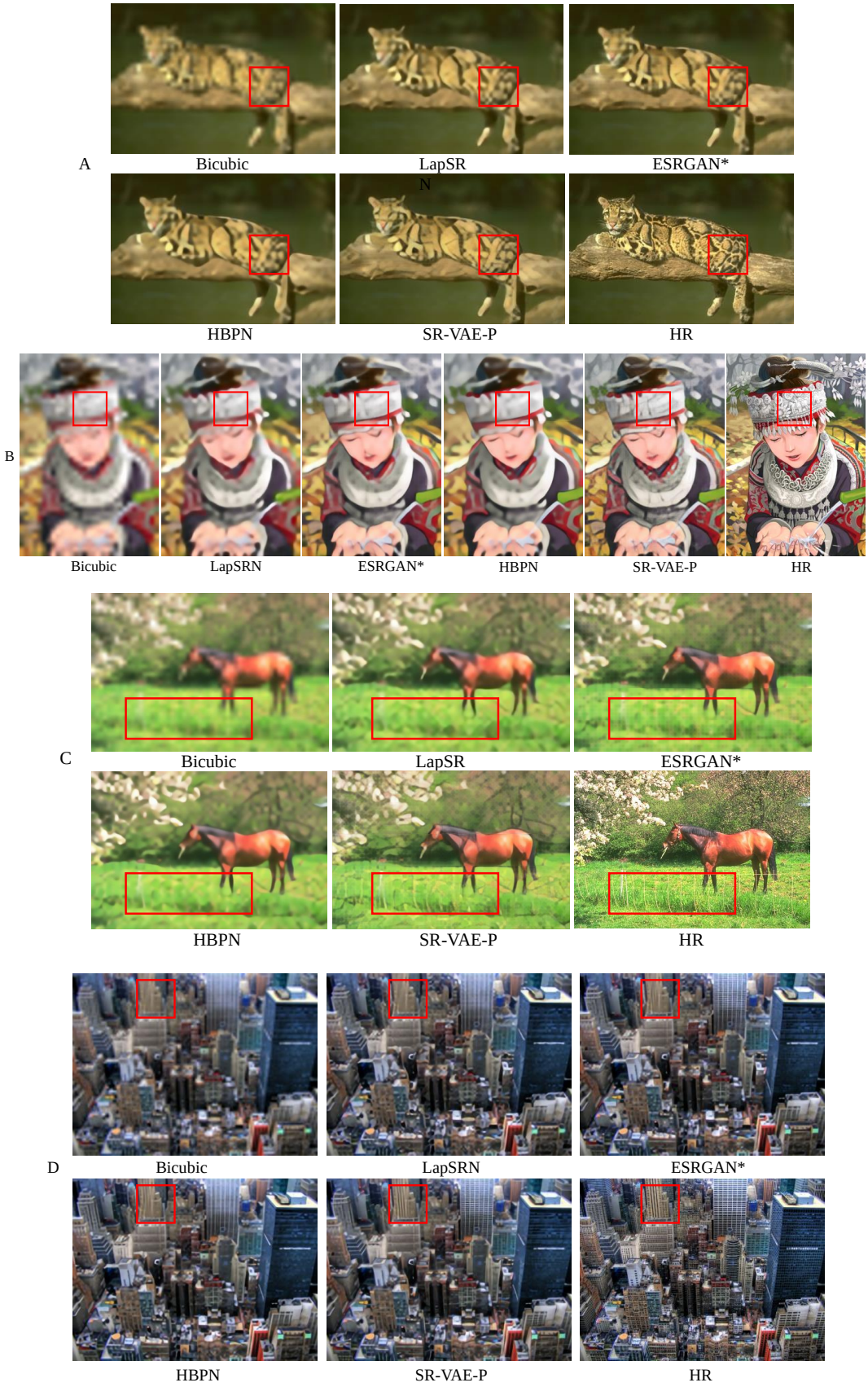

Fig. 13. Visual comparison among different SR algorithms for 8× SR.

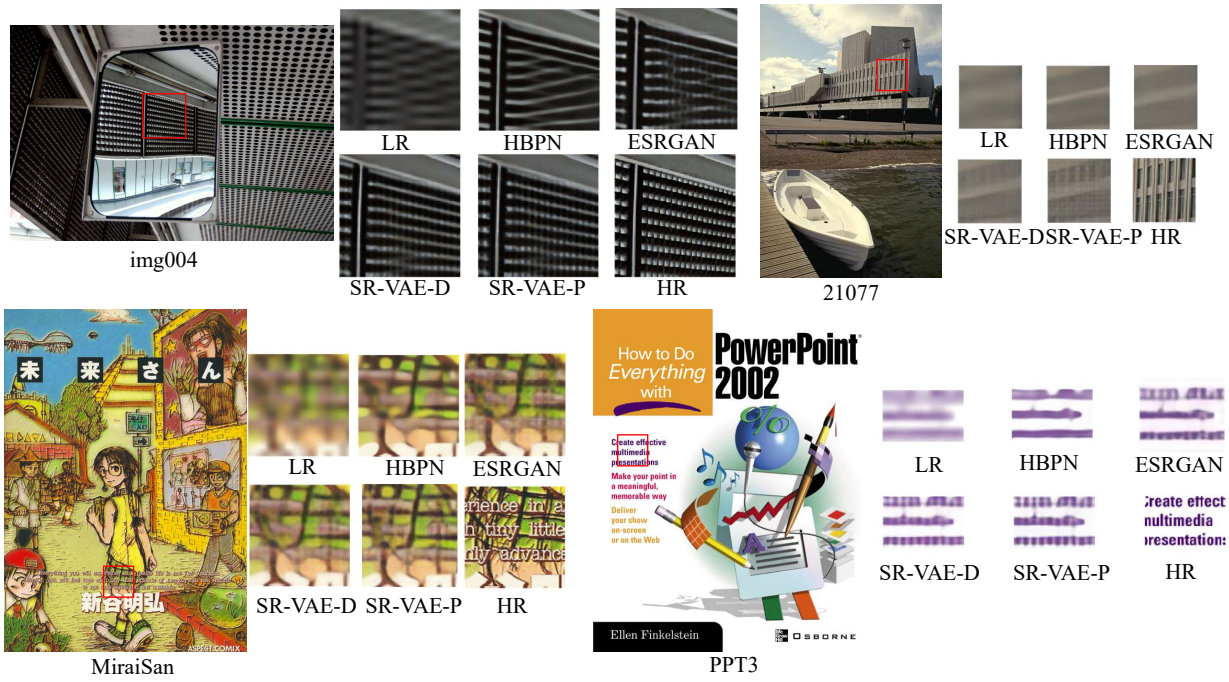

Fig. 14. Visual comparison among different SR algorithms for 8× SR.

SR images are) than other GAN based SR approaches by above 1 dB. It indicates that our approaches balance well between distortion and perception.

- Higher PSNR and better visual quality on 8× image SR. We compare our proposed work to other CNN and GAN based approaches, including the state-of-the-art approaches EDSR, DBPN and HBPN, SRGAN* and ESRGAN*. Note that the authors of SRGAN and ESRGAN did not provide the codes for 8× SR hence we modified the official codes for evaluation. From TABLE IV, we can see that our proposed SR-VAE-D approach (focusing more on distortion) can obtain comparable (even higher) PSNR, SSIM to other state-of-the-art approaches. Meanwhile, the visual comparison in Fig. 13 and 14 also shows that our proposed SR-VAE-P can generate SR results with perceptual quality better than GAN based SR approach. It indicates that our approaches balance well between distortion and perception.

In Fig. 12, we show the results of 4× image SR. Among the SR algorithms, using distortion based approaches (EDSR and HBPN) can generally obtain smooth results (like the hands of clock in image *img002* and the letters in image *21077*) while using GAN based approaches (EnhanceNet, SRGAN, ESRGAN and proposed SR-VAE) can generate SR images with sharp features. For example, image *woman* from Set5, though EnhanceNet and SRGAN gives lower PI values from TABLE IV, they cannot properly reconstruct the right eye of the woman. Using the proposed SR-VAE-P can recover the eye with better visual quality. For image *zebra* from Set14, we can observe similar visual results. EnhanceNet, SRGAN and ESRGAN have generated rather "dirty" and messy strips around the leg of the zebra. Using proposed SR-VAE-P can obtain better strip pattern. Furthermore, note that the original HR image has some additional noise around the leg. Using GAN based approaches can recover some fine textures but they do not have the ability to distinguish the noise from the image pattern so that they also enhance the noise pattern. For the proposed SR-VAE-D and SR-VAE-P, they can balance the data fidelity and visual quality. For image *21077* from BSD100 and image *img002* from Urban100, we can also find similar results. It is interesting to find that the clock in *img002*, using proposed SR-VAE-N and SR-VAE-P can see clearly the time, and the second hand.

For 8× image SR, there is no GAN based SR algorithms for comparison. We have used two of the state-of-the-art SR approaches: DBPN and HBPN for visual comparison.

In Fig. 13 and Fig. 14, we show images from different datasets for 8× image SR. In Fig. 13, we show complete images obtained by different SR algorithms and have drawn red boxes to indicate the key differences. We can see that using the proposed SR-VAE-P can achieve photo-realistic SR results with sharp details, like the leopard print pattern of A, crown and background details of B, fence pattern of C and city buildings of D. In Fig. 14, we show detailed comparisons. For image *img004* from Urban100, using the proposed SR-VAE-D and SR-VAE-P can reconstruct the object pattern from the mirror reflection. Both LapSRN and HBPN wrongly predict the pattern as stripes while ESRGAN* cannot generate as sharp pattern as ours. In image *21077*, the window pattern on the building can be recovered by SR-VAE-P while other approaches only generated blurry results. In images *MiraiSan* from Manga109 and *PPT3* from Set14, the contexts can be

better reconstructed by using the proposed SR-VAE-D and SR-VAE-P.

In all visual comparison, we can find good effects on using the proposed SR-VAE for image SR in both $4\times$ and $8\times$ scaling factors. TABLE IV also proves that the proposed SR-VAE can also achieve higher PSNR and SSIM compared to other GAN based image SR algorithms. In terms of both qualitatively and quantitatively, our proposed SR-VAE outperforms other SR algorithms.

V. Conclusion

We have introduced our proposed Variational AutoEncoders for photo-realistic image super-resolution. Unlike the previous methods which use either CNN based approaches for distortion minimization or GAN based approaches for perception reconstruction, our proposed SR-VAE is the first work that makes use of variational autoencoders for image SR. By introducing conditional sampling generator and SR based encoder and decoder, we can train SR-VAE to learn the conditional generative process on the HR images and also tackle the LR-to-HR mapping relationship for reconstruction simultaneously. The training process can be easily done without using any manual checks and tricks. Furthermore, we have used both distortion and perception based evaluations to measure the performance of the proposed and other works. Comparing with other recent SR algorithms in quantitative and qualitative aspects, SR-VAE can achieve better or state-of-the-art performance. For further study, we will continue to explore the latent parameter optimization via deep learning for image SR. We can design hierarchical variational autoencoders to further extract the conditional distribution of SR and HR images for finer reconstruction. Due to the "hole" problem of the marginal posterior of the training data, more constraints are required to generate sharp and accurate samples. Combining GAN and VAE could form a very promising research direction. Furthermore, other than using PI for measuring visual quality, the evaluation on the perceptual quality can also be further studied to better estimate the visual differences.

Acknowledgment

This work was supported by the Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic university (1-BBA2 and G-YBKG), and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Grant No. PolyU 5243/13E).

References

[1] H. Chang, D.-Y. Yeung, and Y. Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, vol. 1, pp. I–I, Washington DC, USA, June 2004.

[2] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. In *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, March 2002.

[3] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[4] R. Timofte, V. De, and L. V. Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV'13)*, pp. 1920–1927, Sydney, NSW, Dec. 2013.

[5] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Proceedings of the 2014 IEEE Asian Conference on Computer Vision (ACCV'13)*, vol. 9006, pp. 111–126, Singapore, Nov. 2015.

[6] He He and Wan-Chi Siu. Single image super-resolution using gaussian process regression. In *Proceedings of the 2014 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'11)*, pp. 449–456, Providence, RI, June 2011.

[7] Jun-Jie Huang and Wan-Chi Siu. Learning hierarchical decision trees for single-image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 5, pp. 937–950, May 2017.

[8] Jjun-Jie Huang, Wan-Chi Siu, and Tian-Rui Liu. Fast image interpolation via random forests. *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3232–3245, Oct 2015.

[9] F. Humblot and A. Mohammad-Djafari. Super-resolution using hidden markov model and bayesian detection estimation framework. In *Proceedings of the 2006 International Conference on EURASIP Journal on Applied Signal Processing*, 2006, Jan. 2006.

[10] C. Liu, R. Szeliski, S. Bing Kang, C. L. Zitnick, and W. T. Freeman. Automatic estimation and removal of noise from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 299–314, Feb 2008.

[11] Zhi-Song Liu, Wan-Chi Siu, and Yui-Lam Chan. Fast image super-resolution via randomized multi-split forests. In *Proceedings of the 2017 IEEE International Symposium on Circuits and Systems (ISCAS'17)*, pp. 1–4, Baltimore, USA, May 2017.

[12] Zhi-Song Liu, Wan-Chi Siu, and Jun-Jie Huang. Image super-resolution via weighted random forest. In *Proceedings of the 2017 IEEE International Conference on Industrial Technology (ICIT'17)*, pp. 1019–1023, Toronto, Canada, March 2017.

[13] Zhi-Song Liu and Wan-Chi Siu. Cascaded random forests for fast image super-resolution, In *Proceedings of the 2018 IEEE International Conference on Image Processing (ICIP'18)*, pp. 1019–1023, Toronto, Canada, Oct. 2018.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 2015 International Conference on Learning Representations (ICLR'15)*, San Dieago, CA, May, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'16)*, pp. 770–778, Las Vegas, Nevada, June, 2016.

[16] C. Dong, C. Change Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, Feb. 2015.

[17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the 2016 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'16)*, pp. 1646-1654, Las Vegas, Nevada, June, 2016.

[18] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the 2017 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'17)*, Honolulu, Hawaii, June, 2017.

[19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the 2017 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'17)*, Honolulu, Hawaii, June, 2017.

[20] M. Haris, G. Shakhnarovich, and N. Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the 2018 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'18)*, Honolulu, Salt Lake City, Utah, June, 2018.

[21] Zhi-Song Liu, Li-Wen Wang, Chi-Tak Li and Wan-Chi Siu. Hierarchical Back-Projection Networks for image super-resolution. In *Proceedings of the 2019 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'19)*, Long Beach, CA, June, 2019.

[22] U. Schmidt and S. Roth. Shrinkage fields for effective image restoration.

In *Proceedings of the 2014 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'14)*, pp. 2774–2781, Columbus, Ohio, June, 2014.

[23] W. Dong, L. Zhang, G. Shi, and X. Li. Nonlocally centralized sparse representation for image restoration. *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1620–1630, April 2013.

[24] S. Gu, L. Zhang, W. Zuo, and X. Feng. Weighted nuclear norm minimization with application to image denoising. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*, pp. 2862–2869, Columbus, Ohio, June, 2014.

[25] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of the 2014 International Conference on Learning Representations (ICLR'14)*, Calgary, Canada, Dec. 2014.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of the 2014 Advances in Neural Information Processing Systems 27 (NIPS'14)*, pp. 2672–2680, Montreal, Canada, Dec. 2014.

[27] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, pp. 105–114, Honululu, Hawaii, June, 2017.

[28] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks. In *Proceedings of the 2017 International Conference on Learning Representations (ICLR'17)*, Toulon, France, April, 2017.

[29] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, vol. 20, pp. 209–212, 2013.

[30] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *2017 IEEE Conference on Computer Vision (ICCV'17)*, pp. 4491–4500, Venice, Italy, Oct. 2017.

[31] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 2016 Advances in Neural Information Processing Systems (NIPS'16)*, Barcelona, Spain, Dec. 2016.

[32] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *Proceedings of the 2016 International Conference on Learning Representations (ICLR'16)*, San Juan, Puerto Rico, April, 2016.

[33] R. Timofte, S. Gu, J. Wu, and L. Van Gool. Ntire 2018 challenge on single image super-resolution: Methods and results. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'18)*, Salt Lake City, Utah, June, 2018.

[34] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. 2018 PIRM challenge on perceptual image super-resolution. In *Proceedings of the 2018 IEEE European Conference on Computer Vision Workshops (ECCVW'18)*, Munich, Germany, Sep. 2018.

[35] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang. ESRGAN: enhanced super-resolution generative adversarial networks. In *Proceedings of the 2018 IEEE European Conference on Computer Vision Workshops (ECCVW'18)*, Munich, Germany, Sep. 2018.

[36] C. Ma, C. Yang, X. Yang and M. Yang. Learning a no-reference quality metric for single-image super-resolution. *Journal of Computer Vision and Image Understanding*. pp. 1–16, vol. 158, May, 2017

[37] Y. Blau and T. Michaeli. The Perception-Distortion Tradeoff. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. pp. 6228–6237, Salt Lake City, Utah, June, 2018

[38] X. Hou, L. Shen and G. Qiu. Deep Feature Consistent Variational Autoencoder. In *Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV'17)*. pp. 1133–1141, CA, USA, Mar. 2017

[39] J. Engel, M. Hoffman and A. Roberts. Latent Constraints: Learning to Generate Conditionally from Unconditional Generative Models. In *Proceedings of the 2016 International Conference on Learning Representations (ICLR'18)*. Vancouver, Canada, April 2018

[40] A. Makhzani, J. Shlens, N. Jaitly and I. Goodfellow. Adversarial Autoencoders. In *Proceedings of the 2016 International Conference on Learning Representations (ICLR'16)*, San Juan, Puerto Rico, May 2016

[41] S. Kihyuk, L. Honglak and Yan Xinchen. Learning Structured Output Representation using Deep Conditional Generative Models. In *2015 International Conference on Advances in Neural Information Processing Systems (NIPS'15)*. pp. 3483–3491, Montreal, Canada, Dec. 2015

[42] D. Kingma, T. Salimans and M. Welling. Improving Variational Inference with Inverse Autoregressive Flow. In *Proceedings of the 2016 International Conference on Learning Representations (ICLR'16)*. San Juan, Puerto Rico, May 2016

[43] D. Kingma, T. Salimans and M. Welling. Variational Dropout and the Local Reparameterization Trick. In *2015 International Conference on Advances in Neural Information Processing Systems (NIPS'15)*. pp. 3483–3491, Montreal, Canada, Dec. 2015

[44] G.E. Hinton and R.R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. In *Science (New York, N.Y.)*. pp. 504-7, vol. 313, August 2006

[45] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez and A. Courville. PixelVAE: A Latent variable model for natural images. In *International Conference on Learning Representations (ICLR'17)*. Toulon, France, Apr. 2017

[46] A. V. D. Oord, O. Vinyals and K. Kavukcuoglu. Neural Discrete Representation Learning. In *31st Conference on Neural information Proceedings Systems (NIPS'17)*. Long Beach, CA, USA, Dec. 2017

[47] Y. Zhang, K. Li, K Li, L Wang, B Zhong, and Y Fu. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *IEEE on European Conference on Computer Vision (ECCV'2018)*. Munich, Germany, Sep. 2018.

[48] M. Irani and S. Peleg. Motion Analysis for Image Enhancement: Resolution, Occlusion, and Transparency. *Journal of Visual Communication and Image Representation*. pp. 324–335, vol. 4 1993.

[49] A. C. Bovik, H. R. Sheikh and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*. pp. 600–612, vol. 13, No. 4, April 2004.

[50] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 33, no. 5, pp.898–916, May 2011.

[51] M. Bevilacqua, A. Roumy, C. Guillemot, Marie-Line and Alberi Morel. Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In *Proceedings of the 2012 International British Machine Vision Conference (BMVC'12)*, Guildford, Surrey, United Kingdom, Sept. 2012.

[52] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *Proceedings of the 7th International Conference on Curves and Surfaces*, pp. 711–730, Berlin, Heidelberg, 2012. Springer-Verlag.

[53] J. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*, pp. 5197–5206, June 2015.

[54] Y. Matsui, K. Ito, Y. Aramaki, T. Yamasaki, and K. Aizawa. Sketch-based manga retrieval using manga109 dataset. In *Proceedings of the 2017 Multimedia Tools and Applications (MTAP)*, Springer, 2017.

[55] K. He, X. Zhang, S. Ren and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15)*, Santiagao, Chile, pp. 1026–1034, Dec. 2015.

Zhi-Song Liu received the MSc degree in electronic engineering, in 2015, from The Hong Kong Polytechnic University, Hong Kong, where he is currently working toward the PhD degree under the supervision of Prof. Wan-Chi Siu and Dr. Yui-Lam Chan. His research interests include deep learning techniques, image and video signal processing, image and video super-resolution.

Wan-Chi Siu (S'77-M'77-SM'90-F'12-Life-F'16) received the MPhil and PhD degrees from The Chinese University of Hong Kong in 1977 and Imperial College London in 1984. He is a CEng, Fellow of IET and Life-Fellow of IEEE, and has become Emeritus Professor in Department of Electronic and Information Engineering (EIE), The Hong Kong Polytechnic University since 2017. Prof. Siu was Chair Professor between 1992 and 2017, founding Director of the Centre for Signal Processing, Head (EIE) and subsequently Dean of Engineering Faculty between 1994 and 2002 of the same university. He was an independent non-executive director (2000-2015) of a publicly-listed video surveillance company in Hong Kong. Professor Siu is an expert in digital signal processing and machine learning, specializing in fast algorithms, conventional learning, deep learning, video coding, object/pattern recognition and super-resolution imaging for visual surveillance, autonomous vehicle and smart city applications. He has published over 500 research papers (with 200 appeared in international journals such as IEEE Transactions on Image Processing), filed (granted) recently 8 patents, and edited three books. His works on motion estimation, DCT, transcoding, super-resolution imaging, face/object recognition algorithms, etc. are well received by academic peers with high citations, and many of his research works have also been ported into industrial uses for contributions in hi-tech development. Prof. Siu is now Immediate Past-President (2019-2020), and was President (2017-2018) of the Asia-Pacific Signal and Information Processing Association (APSIPA). He was a Vice President of the IEEE Signal Processing Society (2012-2014), and Chairman of Conference Board and a core member of the Board of Governors. Prof. Siu has been Subject Editor, Guest Editor and Associate Editor of a number journals, including recently as Subject Editor (2015-2018, in charge of Image Processing) of Electronics Letters, Associate Editor (2015-2017) of EEE Transactions on Circuits & Systems for Video Technology, and Associate Editor (2012-2014) of IEEE Transactions on Image Processing. He is a very popular lecturing staff member within the University, while outside the University he has been a keynote speaker of over 10 international/national conferences in the recent 10 years. He received many awards, such as Distinguished Presenter Award, the Best Teacher Award, Best Faculty Researcher Award (two times) and the IEEE Third Millennium Medal. He was the General Chair/Technical Program Chair of several prestigious IEEE Society sponsored flagship international conferences (including the ICIP'2010, ICASSP'2003 and ISCAS'1997). He was the chairman of many assessment panels and committees of professional bodies, including as the convenor of the first Engineering and Information Technology Panel of the Research Assessment Exercise (RAE) in Hong Kong, 2003/4. He is now a member of the IEEE Fourier Award for Signal Processing Committee (2018-2020), in additional to some other IEEE Technical Committees.

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 actively involved in professional activities. He has authored over 110 research papers 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 360-degree omnidirectional video coding. Dr. Chan serves as an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING. He was the Secretary of the 2010 IEEE International Conference on Image Processing. He was also 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.