VarSR: Variational Super-Resolution Network for Very Low Resolution Images

Sangeek Hyun[1] and Jae-Pil Heo[1,2]

[1] Dept. of Artificial Intelligence, Sungkyunkwan University
[2] Dept. of Computer Science and Engineering, Sungkyunkwan University

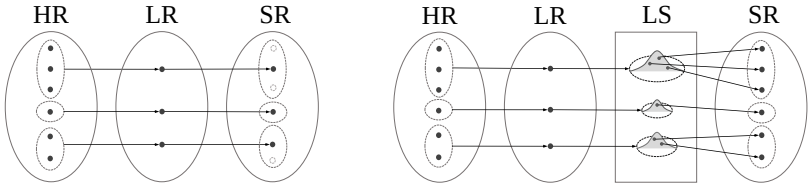
Abstract. As is well known, single image super-resolution (SR) is an ill-posed problem where multiple high resolution (HR) images can be matched to one low resolution (LR) image due to the difference in their representation capabilities. Such many-to-one nature is particularly magnified when super-resolving with large upscaling factors from very low dimensional domains such as 8×8 resolution where detailed information of HR is hardly discovered. Most existing methods are optimized for deterministic generation of SR images under pre-defined objectives such as pixel-level reconstruction and thus limited to the one-to-one correspondence between LR and SR images against the nature. In this paper, we propose VarSR, Variational Super Resolution Network, that matches latent distributions of LR and HR images to recover the missing details. Specifically, we draw samples from the learned common latent distribution of LR and HR to generate diverse SR images as the many-to-one relationship. Experimental results validate that our method can produce more accurate and perceptually plausible SR images from very low resolutions compared to the deterministic techniques.

Keywords: Single image super resolution, variational super resolution, very low resolution image

1 Introduction

Single image super-resolution (SISR) is a fundamental computer vision problem and has a broad range of real-world applications. The recent advances of deep convolutional neural networks (CNNs) have produced great progress in SISR. There has been a lot of CNN architectures to improve the performance of SR in terms of accuracy and time. On the other hand, developing new objectives and quality measures for SR algorithms other than traditional reconstruction errors also has been an active research topic of late years. Notable examples include perceptual similarity, GAN-based loss, and even learned metrics. Thanks to sustained research efforts and recent breakthroughs, the current SR techniques can produce super-resolved images comparable with the original high resolution (HR) ones.

Despite the success of recent SISR techniques, in this paper, we point out the problem of the deterministic mechanism of those methods. SISR is widely known



(a) Deterministic models (b) VarSR-Net

Fig. 1. This figure illustrates the main difference between the deterministic super-resolution models and our proposed VarSR-Net. Although the ill-posed nature of the problem, the deterministic models produce an unique solution so thus they are not able to generate diverse super-resolved results as the high resolution images. However, our proposed VarSR-Net matches the latent distributions of low and high resolution images to produce diverse super-resolved outputs by sampling multiple latent variables from the shared distribution. Note that, 'LS' indicates the matched latent space.

as an ill-posed problem where multiple HR images can share a single matched low resolution (LR) image and thus the super-resolved image is not necessary to be unique. This is mainly caused by the difference in the representation capabilities between LR and HR image spaces. This many-to-one or one-to-many nature of SISR problem is especially magnified when dealing with large upscaling factors from very low resolution images such as 8×8 pixels, since extremely small LR images hardly preserve the detailed information of HR images. However, existing SISR solutions produce a deterministic one-to-one correspondence between LR and SR images against the nature of the problem, since they are mostly optimized for point-to-point error minimization based on strictly paired training samples of LR and its HR ground truth images.

In this paper, we address the aforementioned problem by proposing a novel VarSR-Net that enables us to generate diverse SR images that better reflect the ill-posed nature of the problem. Specifically, we first introduce two different latent variables for LR and HR images which encode their contents, respectively. Two latent variables are learned to have the shared representation so that the HR and LR images can be mapped to a common feature space by minimizing the KL-divergence to bridge the gap between the description capabilities of LR and HR images. Our SR module receives the latent variables as input and it is trained to produce pixel-level accurate and perceptually plausible upscaled images as the ordinary SR techniques. Since the HR images have a higher degree of diversity than LR images, the latent variables of HR images are much denser than LR ones. Thus, we draw multiple samples from the learned common latent distribution to generate diverse SR images in the inference stage. Fig. 1 illustrates the main difference between the deterministic techniques and ours. To our knowledge, it is the first attempt to model and match the latent distributions of LR and HR images and generate diverse SR images from very low resolution LR images.

One might ask why we argue the importance of diverse SR image generation. In theoretical aspect, it is natural that a single LR image is not necessary to

match with an unique HR image as we mentioned. On the other hand, we present the following real-world scenario. Suppose that we have a very low resolution face image of the criminal and we need to super-resolve the LR image before searching in the face image database. In this circumstance, we could miss the chance to identify the criminal if the single result of the deterministic SR model is poor. However, the possibility increases if the SR model produces multiple outputs close to the HR image so that we can perform search many times to make up a sufficient sized shortlist. Similar scenarios can be written such as investigation of the numbers on the very low resolution vehicle license plate images.

Our main contributions are summarized as follows:

- We highlight the problem of deterministic super-resolution approaches which paid less attention to the ill-posed nature of the SR task. Those methods assuming one-to-one correspondence between LR and HR images can fail to recover the detailed information of HR images which is hardly discovered in low dimensional LR images.
- We introduce VarSR, Variational Super-Resolution Network, capable of generating diverse SR results from a single LR image by sampling multiple latent variables from the learned common latent distribution of LR and HR domains.
- Our extensive evaluation with various quality measures validates that our method can produce more accurate and perceptually plausible SR images compared to the deterministic SR techniques.

2 Related Work

2.1 Image super-resolution

Recent deep-learning based single image super-resolution (SISR) techniques can be broadly categorized into two directions, advances in neural network architectures and objective functions.

In aspect of network architectures, Dong et al. [8] successfully developed a three-layered CNN model to SISR and showed superior performance over the handcrafted algorithms. After that, many advanced deep architectures for SR have been proposed. Ledig et al. [18] and Zhang et al. [38] adopt residual blocks and dense blocks toward more accurate SR image reconstruction, respectively. The laplacian pyramid network [16] is proposed to generate multi-scale SR images progressively in one feed-forward pass. Tai et al. [29] reduced the number of parameters of SR networks without loss of depths by using convolutional layers recursively.

In the early stages of deep-learning based SISR, most of the networks [8, 14] are trained under L_p distances as their loss function for pixel-level reconstruction of the HR images. However, models learned for the reconstruction objective alone tend to produce blurry results. To resolve this problem, Johnson et al. [13] proposed the perceptual loss defined by the L_2 distance of activation maps extracted from CNN models pre-trained on large-scale datasets. In addition to the

perceptual loss, SRGAN [18] incorporated the adversarial loss [9] to recover realistic image textures. Furthermore, there have been numbers of SR objectives including texture matching [25], semantic prior [31], and rank loss [37].

To our knowledge, the diversity of SR images is not discussed in the aforementioned previous work. We believe our study can provide a good starting point for the community to have more attention toward the new research direction.

2.2 Super-resolving very low resolution images

The SR methods specialized for very low resolution images have mostly focused on human face images with 8×8 or 16×16 pixels [35, 12, 34, 40, 5]. Yu et al. [33] exploited facial attributes to train SR networks based on conditional GAN [22]. The structural prior of face such as facial landmarks is actively utilized for super-resolving face images [4, 3, 32]. There also have been several attempts to utilize person identities as a constraint when training SR models [10, 7]. Note that, the aforementioned recent techniques tailored for the face image SR exploit additional supervision while our VarSR network is trained totally unsupervised manner.

On the other hand, Dahl et al. [5] proposed recursive learning that generates SR images pixel-by-pixel based on the autoregressive model [23]. Their fully probabilistic model generates diverse results, however, it suffers from the expensive sampling costs proportional to the image size and the number of inferences. On the contrary, our VarSR network can generate diverse and plausible images in one feed-forward path. Moreover, their objective solely concentrates to generate realistic images while our method focuses on the accurate reconstruction of original images.

2.3 Multimodal generative models

Deep-learning models produce a deterministic output unless there are no components for stochasticity. Therefore, there has been a series of work injecting stochasticity into conditional generative models (e.g. adding noise to input) in various tasks. For instance, VAE-based techniques to model the uncertainty were proposed and achieved the state-of-the-art performance in the video prediction task [2, 6]. The BicycleGAN [39] encodes the styles of images as low dimensional latent variables and utilizes randomly sampled latent vectors in the image-to-image translation task. Lee et al. [19] proposed the DRIT by extending BicycleGAN for unpaired image-to-image translation based on a disentangled representation. Unlike the aforementioned methods for improving multimodality in the video prediction or image-to-image translation tasks, our method is the first attempt to introduce latent variables toward diversity in the image super-resolution task.

3 Our Approach

3.1 Motivation

In this paper, we mainly focus on generating multiple SR images with diversity for a single LR image. Before introducing our proposed solution, we clarify our motivation:

Single image super resolution is an ill-posed problem, since a single LR image and multiple HR images can correspond. The one-to-many relationship is basically coming from the difference in the representation capabilities of two domains. The information of HR images cannot be retained by the LR representation without any loss, thus perfect estimation of HR images is inherently impossible. However, most existing SR techniques are optimized to produce deterministic outputs regardless of the nature of the problem.

The problem is especially magnified when super-resolving with high upscaling factors from very low resolution images such as 8×8 pixels where it is extremely hard to discover the high-frequency details of HR images. In this circumstance, it is highly probable that several HR images can be matched to a single LR image. Therefore, a super-resolved image is not necessary to be unique but should be diverse.

There are critical real-world applications including surveillance systems that inferring diverse SR results is beneficial. One practical scenario is described in Sec. 1. In those applications, it is preferred to have a set of candidate SR images highly likely to contain an element very close to the target HR image, rather than a single deterministic result which moderately recovers the HR image.

3.2 Variational Super-Resolution Network

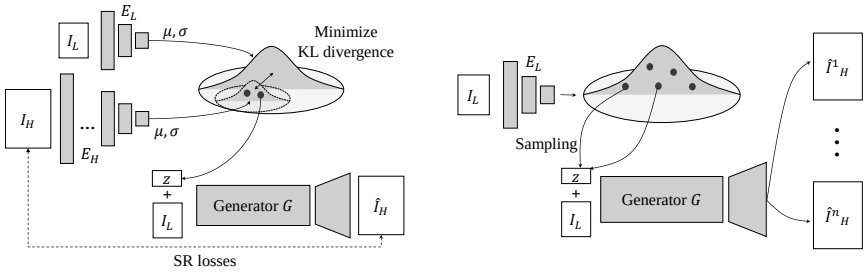
According to our motivation, we propose Variational Super-Resolution Network (VarSR-Net), specialized for the SR tasks from very low resolution images such as 8×8 dimensions where the details of HR images are hardly noticeable, by matching the latent distributions of LR and HR images.

A super-resolution network g_{SR} is a generator to produce a single output image $\hat{I}_H$ for a given low resolution image I_L:

$$\hat{I}_H = g_{SR}(I_L). \tag{1}$$

The network is trained to minimize the distortion (i.e. reconstruction error) between a super-resolved image $\hat{I}_H$ and a corresponding ground-truth high resolution image I_H. Additional objectives such as perceptual similarities or discrimination scores based on GAN can be assigned to produce better imitations. However, the missing detailed information of I_H hardly encoded in I_L makes the super-resolution models fail to infer reliable outputs, especially for very low resolution input.

In order to address this problem, a simple choice is to give an additional latent variable to our SR sub-module g_{SR} that describes the information of I_H



(a) Training (b) Inference

Fig. 2. This figure illustrates the training and inference procedures of VarSR-Net. The HR and LR encoders are stochastically modeled to output parametric distributions. Specifically, encoders estimate a multivariate Gaussian distribution $\mathcal{N}(\mu, \sigma)$ where the their input highly likely to belong to. (a) The E_H and G are trained as an encoder-decoder structure. Specifically, given a pair of training sample (I_L, I_H), the E_H is trained to extract features of I_H while the generator $G(I_L, z)$ is learned to reconstruct I_H from I_L and z sampled from $\mathcal{N}(\mu, \sigma)$ estimated by E_H. Meanwhile, the E_L is learned to minimize the KL divergence with E_H to match the latent distributions of LR and HR images. (b) In inference stage, E_L estimates the latent distribution $\mathcal{N}(\mu, \sigma)$ from the input LR image. Diverse SR images are then produced by $G(I_L, z_i)$ where z_i are randomly drawn latent variables from $\mathcal{N}(\mu, \sigma)$ predicted by E_L.

to resolve the ambiguity that I_L has:

$$\hat{I}_H = g_{SR}\Big(I_L, E_H(I_H)\Big), \tag{2}$$

where $E_H(\cdot)$ is an encoder to extract features from the high resolution images. The SR model g_{SR} then learns to super-resolve a low resolution image I_L much more accurately by exploiting the feature extracted from I_H, since the features $E_H(I_H)$ provide the missing but strong cues for recovering HR images. Furthermore, joint learning of g_{SR} and $E_H(I_H)$ encourages the feature extractor to focus more on the information complementary to I_L.

However, the aforementioned approach (Eq. 2) is totally contradictory since the generator requires I_H or its features $E_H(I_H)$ to produce I_H. In order to resolve such a circular logic, we need to estimate $E_H(\cdot)$ only based on a low resolution image I_L.

We introduce another encoder $E_L(\cdot)$ for low resolution images. As our motivation and the ill-posed nature of super-resolution problem, the feature extracted from a single LR image needs to be matched to several features of HR images. However, it cannot be fulfilled by deterministic encoders. Therefore, we model the latent representation of both encoders $E_H(\cdot)$ and $E_L(\cdot)$ as multivariate Gaussian distributions where we can sample multiple latent variables to provide one-to-many mappings as follows:

$$E_L(x) = [\mu_x, \sigma_x] \text{ and } E_H(x) = [\mu_x, \sigma_x], \tag{3}$$

where x is either of LR and HR image, and both μ_x and σ_x are D-dimensional vector. Note that, D is the dimensionality of the latent representation. We also denote $E_{\cdot}^{\mu}(x) = \mu_x$ and $E_{\cdot}^{\sigma}(x) = \sigma_x$ for the sake of simplicity.

Our key idea is to match the two latent distributions of $E_H(\cdot)$ and $E_L(\cdot)$. In other words, we aim to realize that the sampled latent variables from E_L are highly likely to be ones sampled from E_H. To this end, we train both encoders to minimize the KL divergence between two distributions of $E_H(I_H)$ and $E_L(I_L)$. We further explain more details about the KL divergence term later.

In training phase, the input to our generator $G(\cdot,\cdot)$ is pairs of (I_L, z) where z is a sampled variable from $\mathcal{N}(E_H^{\mu}(I_H), E_H^{\sigma}(I_H))$, since the HR images are available. On the other hand, in testing phase, the generator receives a pair of (I_L, z) where z is drawn from $\mathcal{N}(E_L^{\mu}(I_L), E_L^{\sigma}(I_L))$. Specifically, the n diverse SR images $\{\hat{I}_H^1, ..., \hat{I}_H^n\}$ super-resolved from a LR image I_L are obtained as the follows:

$$\hat{I}_H^i = G(I_L, z_i), \qquad z_i \sim \mathcal{N}\Big(E_L^{\mu}(I_L), E_L^{\sigma}(I_L)\Big), \tag{4}$$

for $i \in \{1, ..., n\}$. Although we utilize the latent distribution predicted by the LR encoder E_L in the inference stage, our generator $G(\cdot,\cdot)$ is capable to recover the information of HR images since E_L is trained to share the common latent distribution with the HR encoder E_H.

Relation with CVAE. Conditional Variational AutoEncoder (CVAE) [27] approximates the conditional distribution $p_\theta(x|y)$ where x is data and y is a condition. The conditional generative process of the model is as follows; for a given condition y, latent variable z is drawn from the prior distribution $p_\theta(z|y)$, and the output x is generated from the distribution $p_\theta(x|y,z)$. This process allows to generate diverse outputs $\{x_i\}$ through the sampling of multiple latent variables $\{z_i\}$. Variation lower bound for CVAE is defined as:

$$\begin{aligned} L_{CVAE}(x, y; \theta, \phi) = \mathbb{E}_{q_\phi(z|x,y)} \log p_\theta(x|y,z) \\ - D_{KL}(q_\phi(z|x,y) \| p_\theta(z|y)) \leq \log p_\theta(x|y), \end{aligned} \tag{5}$$

where $q_\phi(z|x,y)$ is the approximated distribution of the true posterior and D_{KL} is the KL divergence.

If we assume that a high resolution image contains all the information of its low resolution counterpart, we can translate our VarSR-Net network to CVAE architecture; x as high resolution image I_H, y as low resolution image I_L, $p_\theta(z|y)$ as a LR encoder $E_L(I_L)$, $q_\phi(z|x,y)$ as a HR encoder $E_H(I_H)$, and $p_\theta(x|y,z)$ as a generator network $G(I_L, z)$. Also, the term $\log p_\theta(x|y,z)$ can be replaced with losses used in previous SR works such as pixel-level reconstruction or perceptual loss. This interpretation gives a theoretical support for our model that maximizes conditional log-likelihood of observed data.

3.3 Objective functions

We train the entire model in an end-to-end fashion based on a weighted combination of KL divergence and pixel-level losses. The pixel-level reconstruction loss which guides to reduce the distortion of a super-resolved image against its high resolution counterpart encourages the HR encoder $E_H(\cdot)$ to extract the informative features of a high resolution image, while the KL divergence loss minimizes the divergence between latent feature distributions of the high and low resolution images.

Specifically, a pixel loss minimizes the pixel-wise L_2 distances between a ground-truth high resolution image I_H and a super-resolved image. Note that, the super-resolved output is generated with latent variables sampled from a high resolution encoder E_H in training time. Therefore, the pixel-level loss is formulated as follows:

$$\mathcal{L}_{\text{pixel}} = \frac{1}{r^2HW}\sum_{x=1}^{rH}\sum_{y=1}^{rW}\Big(I_H^{x,y} - G(I_L, z)^{x,y}\Big)^2,\ z \sim \mathcal{N}\Big(E_H^{\mu}(I_H), E_H^{\sigma}(I_H)\Big), \quad (6)$$

where r is an upscaling factor, and H, W are height and width of I_L, respectively.

KL divergence loss has the most important role in our framework. It enables the low resolution encoder E_L to infer the latent variables which pretend ones from E_H. The KL divergence loss is formulated as follows:

$$\mathcal{L}_{\text{KL}} = D_{KL}\Big(q(z|I_H) \parallel p(z|I_L)\Big), \quad (7)$$

where $q(z|I_H)$ and $p(z|I_L)$ are the latent feature distributions of $E_H(\cdot)$ and $E_L(\cdot)$, respectively.

Furthermore, we also apply the adversarial loss to recover the realistic texture of high resolution images. We especially adopt the Improved Wasserstein GAN (WGAN-GP) [1, 11] as follows:

$$\mathcal{L}_{\text{adv}} = \mathop{\mathbb{E}}_{\hat{I}\sim\mathbb{P}_g}[D(\hat{I})] - \mathop{\mathbb{E}}_{I\sim\mathbb{P}_r}[D(I)] + \delta \mathop{\mathbb{E}}_{\hat{I}\sim\mathbb{P}_{\hat{I}}}[(\|\nabla_{\hat{I}} D(\hat{I})\|_2 - 1)^2], \quad (8)$$

where D is a critic network. $\mathbb{P}_r$ and $\mathbb{P}_g$ are HR and SR data distributions, respectively. Plus, $\mathbb{P}_{\hat{I}}$ is the distribution of images sampled uniformly along straight lines connecting pairs of points from $\mathbb{P}_r$ and $\mathbb{P}_g$. We use $\delta = 10$ for our experiments.

Finally, our final loss function is given as follows:

$$\mathcal{L} = \lambda_{\text{pixel}}\mathcal{L}_{\text{pixel}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}. \quad (9)$$

where λ_{pixel}, λ_{KL}, and λ_{adv} are hyper-parameters that balance three different loss terms.

4 Experiments

We evaluate our model quantitatively and qualitatively in the human face and digit datasets. Details of each dataset are introduced in the following section. We compare our model against the pixel recursive super resolution (PRSR) [5] which is an auto-regressive model for super-resolution, and MR-GAN [20] that reduces the mode-collapse problem in conditional GAN by replacing the mean squared error (MSE) loss with the momentum reconstruction loss. Unfortunately, PRSR is not tested in the face super-resolution task in our experiments, since it requires very expensive sampling cost for the images of 64×64 pixels. In addition, we utilize SRGAN [18] as the baseline deterministic SR technique for a face super-resolution task. For digit datasets, we use an autoencoder with skip-connections for entire methods since the input low resolution images of digit datasets have extremely low dimensionality such as 2×4 pixels. Therefore, we denote the deterministic baseline as "Det." for digit datasets instead of using SRGAN.

4.1 Datasets

Human face dataset We adopt Celebrity Face Attributes (CelebA) [21] dataset for face super-resolution task. This dataset contains about 200K celebrity facial images. Among them, 100K images are used for training, and other 1K images without any overlap with the training set are utilized as a testing set.

We used a cropped version of CelebA to focus on learning various facial attributes unsupervisingly. We also set a spatial resolution of 64×64 for high resolution image, and 8×8 for low resolution image. For a fair comparison, we allocate 8 residual blocks for all models as did in MR-GAN [20].

Digit datasets For super-resolving digits, we use two datasets: MNIST [17] and license plate(LP) [28] datasets. The MNIST dataset [17] contains handwritten digit images with a resolution of 28×28. There are 60K and 10K images as training and testing sets, respectively. To make the digits unrecognizable in low resolution, all images are downsampled to the resolution of 6×6. The LP dataset [28] is originally designed for the vehicle re-identification task based on low quality LP images labeled in character-level. We collect about 110K and 7K character images cropped from the LP images for training and testing, respectively. Each character image is downsampled to a resolution of 2×4. We set the upscaling scale factor to 4 for digit datasets.

4.2 Implementation details

We implement our VarSR-Net based on the architectures of U-net [24] and SR-GAN [18]. Further architectural details are available in supplementary materials. We set the dimension for latent representation to 256 and 64 for face and digits datasets, respectively. Adam optimizer [15] is utilized with $\beta_1 = 0.9$, $\beta_2 = 0.99$

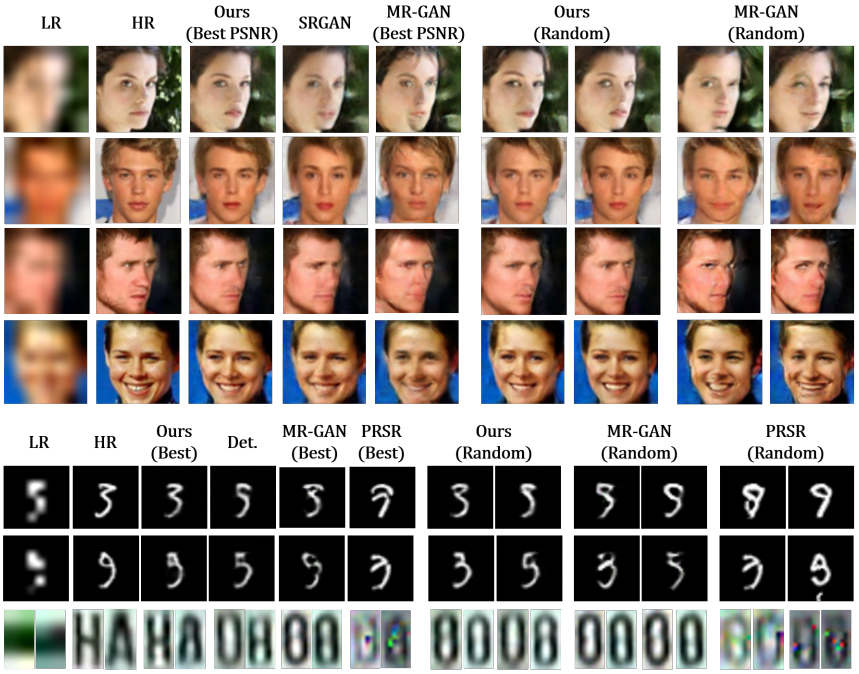
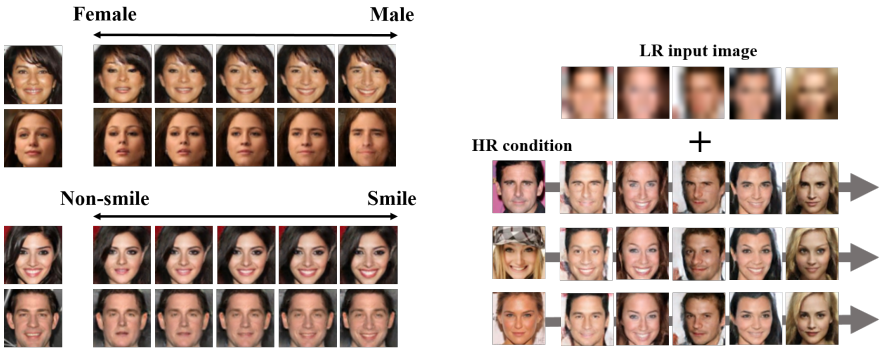


Fig. 3. Qualitative comparison of different super-resolution methods. Images with the highest PSNR scores among stochastically generated diverse results are reported with '(Best PSNR)' or '(Best)'. Randomly sampled images from stochastic models are reported with '(Random)'.

and set the learning rate 1e-4 which is decayed once throughout entire training. As SRGAN [18] did, we first train the model without the adversarial loss to avoid to falling into undesirable local minima. We borrow the architecture of SRGAN's discriminator. However, we remove 3 convolution layers from the discriminator for digit datasets. For normalization layer, we adopt Instance Normalization [30]. Our hyper-parameters to balance the magnitudes of each terms in loss function (Eq. 9) are set as follows; $\lambda_{\text{pixel}} = 1$, and $\lambda_{\text{adv}} = 0.001$ for all datasets, and $\lambda_{\text{KL}} = 0.05$ for LP dataset, $\lambda_{\text{KL}} = 0.01$ for MNIST dataset, and $\lambda_{\text{KL}} = 0.01$, 0.02 for CelebA dataset. In addition, we adopt perceptual loss [13] to realistic texture of images like SRGAN. We add one subpixel convolution layer before the output layer for SRGAN and MR-GAN to deal with the upscaling scale factor of 8 in the face dataset.

4.3 Evaluation metrics

Since VarSR-Net is not developed to generate a deterministic result, we perform the evaluations based on the mean and best scores among diverse super-resolved images. Traditional image quality measures including PSNR, SSIM, and MSE



(a) Attribute editing (b) Attribute transfer

Fig. 4. Qualitative results in face attribute editing and transfer by latent vector manipulations. (a) Given a latent vector z_{LR} of a LR image I_L downsampled from the leftmost HR image, we generate super-resolved images by $G(I_L, z_k)$ with $z_k = z_{LR} + ks\bar{Z}^{\text{att}}$, where $\bar{Z}^{\text{att}}$ is the mean of latent vectors $E_H(\cdot)$ of HR images having a particular attribute such as "male" or "smile", and s is a scaling constant. The 5 super-resolved images are corresponding to different $k \in \{-2, -1, 0, 1, 2\}$. (b) We denote a pair of LR and HR images, I_L^j and I_H^k, at the j^{th} column on the uppermost row and k^{th} row on the leftmost column, respectively. The images $G(I_L^j, E_H(I_H^k))$ super-resolved from I_L^j with the latent vector $E_H(I_H^k)$ of I_H^k are presented at j^{th} column of k^{th} row of the central 3×5 image array to validate that the attributes of HR images can be transferred to super-resolved images by its latent variable.

are used. In addition, we perform the image classification in digit datasets to quantify how well the models produce semantically reliable outputs. For the face dataset, we use the perceptual image quality metrics to quantitatively measure the capabilities of tested methods to generate perceptually plausible images. Specifically, we utilize LPIPS score [36] and the distance between features extracted by a face verification network FaceNet [26]. In addition, we also measure the diversity of super-resolved face images as the average LPIPS distance among multiple resulting images according to Zhu et al. [39].

4.4 Qualitative results

The qualitative results on the human face and digit datasets are shown in Fig. 3. The deterministic baseline suffers from the blurriness of outputs due to the inherited uncertainty of one-to-many mapping and generates inappropriate attributes. Besides, MR-GAN and PRSR succeed to generate diverse outputs, however, the most resulting images are perceptually unsatisfied while our VarSR-Net produces diverse and visually realistic images.

Furthermore, we validate that our common latent distribution shared by LR and HR domains reflects high-level semantics. We specifically perform attribute editing and transfer via manipulation of latent vectors, and the results

Table 1. Quantitative results on CelebA dataset. For Ours and MR-GAN, the PSNR, SSIM, and MSE scores are computed with the image having the best PSNR among 10 samples, while the best LPIPS and Facenet scores within 10 samples are reported.

	w/o adversarial loss			w/ adversarial loss			
	SRResNet	Ours (λ_{KL}=1e-2)	Ours (λ_{KL}=2e-2)	SRGAN	MR-GAN	Ours (λ_{KL}=1e-2)	Ours (λ_{KL}=2e-2)
PSNR	22.57	22.46	**22.74**	22.28	21.21	21.87	22.14
SSIM	0.7242	0.7162	**0.7278**	0.7042	0.6476	0.6855	0.6948
MSE	73.93	72.87	**72.30**	74.49	77.77	74.64	73.74
LPIPS	0.1172	0.0885	0.0927	0.0679	0.0591	0.0539	0.**0538**
Facenet	0.0463	0.0425	0.0430	0.0463	0.0434	**0.0422**	0.0426

Table 2. Diversity and consistency measure on CelebA dataset.

	SRGAN	MR-GAN	Ours (λ_{KL}=1e-2)	Ours (λ_{KL}=2e-2)
Diversity	0.0000	**0.0665**	0.0353	0.0238
Consistency	3.595	4.139	**3.402**	3.411

are shown in Fig. 4. We observe that the edited results successfully reflect the attributes encoded by the conditional latent vectors. More importantly, edited images maintain the original characteristics of the given input LR images. This confirms that the latent vectors estimated by encoders describe the high-level semantics that is complementary to the information LR images have.

4.5 Quantitative results

We perform the quantitative experiments on the human face dataset. We generate 10 SR images for each LR image input. The tested methods are categorized into two groups depending on; training with adversarial loss or not. The results of traditional metrics are shown in Table 1. Without adversarial loss, our model with λ_{KL}=2e-2 achieves the highest scores in PSNR/SSIM/MSE, and the model with λ_{KL}=1e-2 also shows comparable results to the baselines. Unlike the aforementioned cases, SRGAN shows better performance than ours when trained with adversarial loss. However, our models are still significantly better than another stochastic model MR-GAN and accomplished higher scores in perceptual metrics.

In order to measure diversity, we compute the average LPIPS scores among generated samples. Furthermore, we also evaluate the consistency of generated images by measuring the pixel-level L_1 distance between an input low resolution image and the downsampled version of generated SR images. A higher consistency score indicates that resulting SRs are less relevant to the corresponding HR image. Note that, the same downsampling scheme to construct the training set is utilized to measure the consistency. As reported in Table 2, we observe that MR-GAN generates more diverse images but shows much lower consistency compared to our method. It is mainly because the MR-GAN generates many

Table 3. Quantitative results on digit datasets. The PSNR, SSIM, and MSE scores of PRSR, MR-GAN, and Ours are computed with the image having the best PSNR among 5 samples. 'Det.' denotes a deterministic model.

	MNIST			LP		
	PSNR	SSIM	MSE	PSNR	SSIM	MSE
Det.	20.64	0.8250	15.81	21.16	0.9354	88.50
PRSR	18.04	0.7637	15.62	19.22	0.8793	94.45
MR-GAN	21.05	0.8512	15.19	21.54	0.9422	87.93
Ours	**22.00**	**0.8642**	**14.97**	**22.00**	**0.9494**	**85.95**

Table 4. Classification results of super-resolved images in MNIST and LP datasets. "Best" is measured by the criteria that there is at least one correctly classified image among 5 sampled images, and "Mean" is the average softmax values of 5 samples. Note that, the classification accuracies for ground truth HR images are 98.04% and 99.24% for MNIST and LP datasets, respectively.

	MNIST		LP	
	Best	Mean	Best	Mean
Det.	84.74	84.74	93.14	**93.14**
PRSR	84.42	70.22	96.10	92.19
MR-GAN	92.73	84.44	95.53	91.87
Ours	**92.79**	**85.58**	**96.33**	93.01

samples less relevant to the HR ground truth. On the other hand, our method shows even higher consistency score than the deterministic model.

We now perform the quantitative experiments on digit datasets. We sample 5 images for each low resolution image in digit datasets. In Table 3, our model shows superior performance in traditional metrics in both datasets. Also, we observe that the proposed model achieves higher classification accuracy compared to baseline models as reported in Table 4. Those results support that the deterministic model generates semantically or visually incorrect images in both MNIST and LP datasets. On the other hand, MR-GAN generates diverse outputs that contain correctly classified images, but shows low scores in distortion measures as shown in Table 3. PRSR records the lowest score except for classification results in the LP dataset. Note that, a low score of "Mean" in the tables do not degrade our method, because the proposed model can suggest the correct outputs which indicated by the "Best" in Table 4.

One may argue that the "Best" accuracy in Table 4 can be unfair because higher score can be achieved if the model produce just diverse digits regardless of the input. For instance, if a model always produces 5 different digits, the accuracy of "Best" must be largely improved. Therefore, we measure the number of distinct classes within 5 predictions and report in Table. 5. In most cases, 5 sampled images belong to one or two classes. This validates that our model does not just generate diverse images but produce semantically reasonable results.

Table 5. Semantic consistency. The number of distinct classes within 5 generated samples in digit datasets is measured to evaluate semantic consistency. Note that, the class labels are predicted by a pre-trained digit classification model. In more than 90% of cases, the 5 samples produced by our method belong to less than 2 different semantic classes. The scores of "Best" are measured by a criteria that there is at least one correctly classified image among 5 sampled images.

		The number of distinct classes among 5 samples				
		1	2	3	4	5
MNIST	Ratio	73.52%	20.18%	4.86%	0.83%	0.06%
	Best	94.83	88.02	82.95	82.18	87.96
LP	Ratio	89.29%	9.67%	0.94%	0.08%	0.00%
	Best	96.85	92.24	89.92	96.66	-

5 Conclusion

In this paper, we have highlighted the ill-posed nature of the single image super-resolution (SR) problem where multiple high resolution (HR) images can have a common matched low resolution (LR) image due to the difference in their representation capabilities. Despite of the many-to-one nature of the problem, most previous super-resolution techniques deterministically generate outputs. To establish the diversity of super-resolved images, we have modeled stochastic latent distributions for both LR and HR domains and train our VarSR-Net to match two distributions by adopting the KL divergence. The LR encoder learns to imitate the latent distribution of the HR encoder while the HR encoder is trained to extract the informative features of HR images. To produce diverse super-resolved images, we sample multiple latent variables from the parametric distribution estimated by the LR encoder. We intensively evaluate our proposed VarSR-Net against deterministic SR models, and the experimental results validate that our method is capable to produce more accurate and perceptually plausible SR images from very low resolution images. To our knowledge, our VarSR-Net is the first stochastic attempt to overcome the underdetermined characteristic of the SR problem. We believe that our work will stimulate the researcher to pay more attention to resolve the ill-posed nature of the SR problem.

Acknowledgements

This work was supported in part by Samsung Research Funding & Incubation Center for Future Technology (SRFC-IT1901-01), Police Lab (NRF-2018M3E2A1081572), and AI Graduate School Support Program (MSIT/IITP 2019-0-00421).
Jae-Pil Heo is the corresponding author of this paper.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)
2. Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R.H., Levine, S.: Stochastic variational video prediction. arXiv preprint arXiv:1710.11252 (2017)
3. Bulat, A., Tzimiropoulos, G.: Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 109–117 (2018)
4. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2492–2501 (2018)
5. Dahl, R., Norouzi, M., Shlens, J.: Pixel recursive super resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5439–5448 (2017)
6. Denton, E., Fergus, R.: Stochastic video generation with a learned prior. arXiv preprint arXiv:1802.07687 (2018)
7. Dogan, B., Gu, S., Timofte, R.: Exemplar guided face image super-resolution without facial landmarks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
8. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
10. Grm, K., Scheirer, W.J., Štruc, V.: Face hallucination using cascaded super-resolution and identity priors. IEEE Transactions on Image Processing **29**(1), 2150–2165 (2019)
11. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in neural information processing systems. pp. 5767–5777 (2017)
12. Huang, H., He, R., Sun, Z., Tan, T.: Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1689–1697 (2017)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
14. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 624–632 (2017)
17. LeCun, Y., Cortes, C.: MNIST handwritten digit database (2010), http://yann.lecun.com/exdb/mnist/

18. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
19. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European conference on computer vision (ECCV). pp. 35–51 (2018)
20. Lee, S., Ha, J., Kim, G.: Harmonizing maximum likelihood with gans for multimodal conditional generation. arXiv preprint arXiv:1902.09225 (2019)
21. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
22. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
23. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. In: Advances in neural information processing systems. pp. 4790–4798 (2016)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
25. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4491–4500 (2017)
26. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
27. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in neural information processing systems. pp. 3483–3491 (2015)
28. Špaňhel, J., Sochor, J., Juránek, R., Herout, A., Maršík, L., Zemčík, P.: Holistic recognition of low quality license plates by cnn using track annotated data. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6. IEEE (Aug 2017). https://doi.org/10.1109/AVSS.2017.8078501
29. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3147–3155 (2017)
30. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
31. Wang, X., Yu, K., Dong, C., Change Loy, C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 606–615 (2018)
32. Yu, X., Fernando, B., Ghanem, B., Porikli, F., Hartley, R.: Face super-resolution guided by facial component heatmaps. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 217–233 (2018)
33. Yu, X., Fernando, B., Hartley, R., Porikli, F.: Super-resolving very low-resolution face images with supplementary attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 908–917 (2018)
34. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: European conference on computer vision. pp. 318–333. Springer (2016)

35. Yu, X., Porikli, F.: Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3760–3768 (2017)
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
37. Zhang, W., Liu, Y., Dong, C., Qiao, Y.: Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3096–3105 (2019)
38. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2472–2481 (2018)
39. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in neural information processing systems. pp. 465–476 (2017)
40. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: European conference on computer vision. pp. 614–630. Springer (2016)