Variational AutoEncoder for Reference based Image Super-Resolution

Zhi-Song Liu[1]

[1]Caritas Institute of Higher Education

zsliu@cihe.edu.hk

Wan-Chi Siu[1,2]

[1,2]The Hong Kong Polytechnic University

enwcsiu@polyu.edu.hk, liwen.wang@connect.polyu.hk

Li-Wen Wang [2]

Abstract

*In this paper, we propose a novel reference based image super-resolution approach via Variational AutoEncoder (*RefVAE*). Existing state-of-the-art methods mainly focus on single image super-resolution which cannot perform well on large upsampling factors, e.g., 8×. We propose a reference based image super-resolution, for which any arbitrary image can act as a reference for super-resolution. Even using random map or low-resolution image itself, the proposed RefVAE can transfer the knowledge from the reference to the super-resolved images. Depending upon different references, the proposed method can generate different versions of super-resolved images from a hidden super-resolution space. Besides using different datasets for some standard evaluations with PSNR and SSIM, we also took part in the NTIRE2021 SR Space challenge [29] and have provided results of the randomness evaluation of our approach. Compared to other state-of-the-art methods, our approach achieves higher diverse scores.*

1. Introduction

Image Super-Resolution (SR) is a fundamental problem in image processing. Given a low-resolution (LR) image, the objective is to upsample the LR image by $\alpha\times$ to obtain the SR image. In other words, each LR pixel is used for predicting $\alpha\times\alpha$ HR pixels. In real applications, super-resolution is useful for images/videos display, storage, broadcasting and transmission. In particular, it can provide better visual experience for watching movies and playing video games. [2, 3]

Most works for image super-resolution are limited as they focus on single image super-resolution by optimizing the mean squared errors (MSE) between HR and SR pixels. Due to the ill-posed nature of super-resolution, using MSE usually leads to the blurry effect on edges [6, 20]. One solution is to use Generative Adversarial Network to implicitly align the distribution between SR and HR images. Despite recent development on GAN based image SR [20, 39, 33], they usually focus on smaller upsampling factors, like 4× [39]. For larger upsampling factors, like 16× SR, researchers ei-

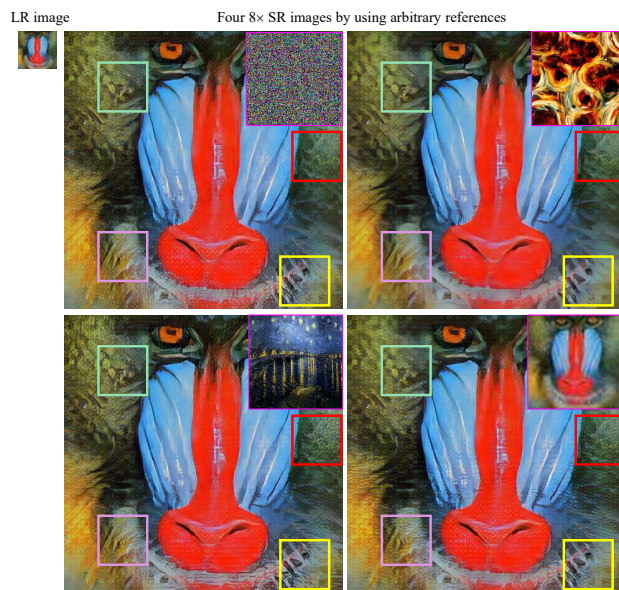Figure 1: **Visualization of 8× SR using our proposed *RefVAE*.** The visualization of the SR images given different references (right upper corner of the SR images), like random noise, random images or LR image itself.

ther use 4× SR model twice, or use simple interpolation to upsample the LR image by 4× first and then use SR model for another 4× upsampling [21]. Our concern is that the single image super-resolution has reached its limit on large upsampling factors even with most advanced deep learning techniques. Hence, we propose a reference based image super-resolution (*RefVAE*), which can search similar patterns from the reference to guide the super-resolution. It is an online searching but fast processing because of the propsoed Conditional Variational Inference (CVAE). It is not an uncommon procedure in image SR using online approach to search similar patches from external [12] or internal [45] datasets for SR. The difference is that we compress various reference images into a compact hidden space by CVAE. Instead of learning complex regression models [15], we propose to use CVAE to learn explicit distribution from reference images, then we sample the patterns from the dis-

tribution as a condition or prior to super-resolve LR image. Different from reference based image super-resolution,hence it is fast for real-time applications. Our proposed method is also different from single image SR. Once the training is done, we can use different "references" to generate different SR candidates, hence, we can expand the SR space to allow users to pick one of the SR candidates for applications. As shown in Figure 1, the "reference" (right-hand upper sub-images) can be any images, even a random Gaussian map or the LR image itself. Our approach can generate a SR image with good visual quality but with different perceptual details around the edges and textures (The differences are marked in boxes with different colors).

To sum up, our contributions include:

- We propose *RefVAE* to explicitly discover image distribution using the proposed Conditional Variational AutoEncoder for image super-resolution.

- Instead of learning directly the data distribution, we introduce references as a condition to guide super-resolution. To transfer the features from style images, we propose to combine both pixel loss and style loss for training. (See Section)

- We provide experimental results on different datasets and show analysis and comparison on both single image SR and reference based image SR results.

2. Related work

In this section, we give a brief review on related previous works on single image super-resolution. To introduce our proposed *RefVAE* method, we also revisit some representative works on generative learning approaches for image SR. Finally, we also introduce related works on reference based image super-resolution.

2.1. Single image super-resolution

Single image super-resolution uses one single LR image to produce the corresponding SR image. It is a classic topic in image processing and a lot of works have been proposed to resolve it. With the development of learning approaches, learning based image SR dominates this field. We can categorize learning based SR into two groups: conventional learning based approaches [36, 23, 22] and deep learning based approaches [28, 43, 21, 27, 18].

Let us focus on some representative deep learning based approaches. Dong et al. [9] proposed the first CNN based image SR using only three layers of convolutions to learn the mapping relations between LR and HR images. Later on, VDSR [18], LapSRN [19], SRResNet [20], EDSR [21] and many other works [43, 44] use a deeper and wider (number of filters) CNN model to learn the mapping functions for super-resolution. Back projection in CNN [13, 27] also sets a new

path for image SR for which iterative residual update can minimize the feature distance for super-resolution. Recent studies [28, 8, 31] show that attention can learn non-local features for super-resolution, especially useful for larger images. Liu et al. [28] proposed an attention based back projection network for large-scale super-resolution. Dai et al. [8] proposed higher-order attention using channel and spatial attention for image. To fully explore the correlations in the channel dimension, Niu et al. [31] proposed a holistic attention network for image super-resolution. The above mentioned approaches train different convolutional neural network end-to-end to optimize mean squared errors between SR and HR pixels. The problem is that using MSE to give equal weights to different pixels can cause blurry effect. To address this problem, many researchers have proposed generative learning approaches for image SR.

2.2. Generative learning approaches for SR

Using generative learning approaches for image SR is a popular topic. We can consider using a single CNN model for SR as discriminative learning that does not explore the data distribution for modelling. On the other hand, generative learning approaches explicitly or implicitly study the data distribution for modelling. Generative Adversarial Network [11] is one of the popular implicit generative approaches. It is also used in image SR. For example, SR-GAN [20] uses GAN to train 4× SR. In order to generate more photo-realistic features, VGG based feature loss is used to minimize the deep feature distances between SR and HR images. ESRGAN [39] proposes to use relativistic GAN [17] to let the discriminator predict relative realness instead of the absolute value. ESRGAN+ [33] further improves ESRGAN by introducing noise for to explore stochastic variations. Vartiaonal AutoEncoder is another choice for SR.

Different from GAN, VAE explicitly explores the data distribution for modelling. Liu et al. [25] first proposed VAE for image super-resolution that achieves comparable performance as CNN based approaches. Furthermore, [24] introduces a reference based face SR that uses the conditional VAE to achieve domain-specific super-resolution. A sampling generator is proposed to constraint the choice of samples for better face reconstruction. To avoid the blurry effect [10] caused by the sampling process in VAE, [26] uses VAE for both image super-resolution and denoising by combining GAN and VAE together so that it can use adversarial loss to encourage photo-realistic feature reconstruction. Most recently, Lugmayr et al. [30] proposed to use conditional normalizing flow to model image distribution for super-resolution and achieved good performance in both quantitative and qualitative results. Normalizing flow based approaches can also be extended to image manipulation [4, 7]. However, the whole normalizing flow structure is over-complex (approx. 50,000,000 parameters) which could

be difficult for training.

2.3. Reference based Super-Resolution (RefSR)

The development of reference based image super-resolution (RefSR) is not a surprise to researchers. Originally, it comes from the non-local filtering. The conventional learning based approaches are patch based process. In other words, the LR patches can be reconstructed by learning the mapping models from images/videos with similar contents [12, 46, 35, 45, 15]. For example, k Nearest Neighbor (kNN) [15] can be used for online search so it limits the big data search. Random Forests [22] is a fast algorithm that can classify training patches into different groups for diverse regression modelling.

Reference based SR can also be achieved by using deep learning. [12] proposes to use an extra HR reference to guide SR. The idea is to use CNN to extract multi-scale HR features to fill the missing information for LR images. Similarly, Zheng et al. [46, 35] proposed a similar reference based SR approach using CNN. The difference is that it uses flow warping operation to align LR and reference features for super-resolution. However, it still requires a pre-aligned reference for SR, or the warping operation cannot well match the feature maps. Most recently, Tan et al. [35] proposed a reference pool to expand the searching region for multi-reference based super-resolution. Not only CNN can be used for RefSR, Liu et al. [24] proposed a face SR that uses an image from the same identity as reference. The idea is to train a conditional Variational AutoEncoder to learn conditional distribution between the reference and LR image. The limitation is that it only investigates the facial image SR and the reference image has to be the face with the same identity. It is not suitable for general image SR.

The study of RefSR also reveals the ill-posed nature of SR that more than one SR image can be formed from a LR image. Depending on different external or internal images as reference, we can transfer reference features to fill out the missing information for LR image enlargement. As described in the Learning the Super-Resolution Space Challenge [29], we can formulate the SR problem as learning a stochastic mapping, capable of sampling from the space of plausible high-resolution images given a low-resolution image. By using an arbitrary reference image to expand data diversity, RefSR is one way to resolve SR space. The advantage is that 1) multiple predictions can be generated and compared and 2) we allow users to choose references with desired patterns for interactive SR.

3. Method

In this section, we will give detailed introduction on the proposed Reference based image super-resolution approach via Variational AutoEncoder (*RefVAE*). Let us formally define the image SR. Mathematically, given a LR image

$\mathbf{X} \in \mathbb{R}^{m \times n \times 3}$ which has been down-sampled from the corresponding HR image $\mathbf{Y} \in \mathbb{R}^{\alpha m \times \alpha n \times 3}$, where (m, n) is the dimension the image and α is the up-sampling factor. They are related by the following degradation model,

$$\mathbf{X} = \mathbf{D} \otimes \mathbf{Y} + \mu \quad (1)$$

where $\otimes$ represents the down-sampling convolution, μ is the additive white Gaussian noise and $\mathbf{D}$ denotes the down-sampling kernel. The goal of image SR is to resolve Equation 1 as the Maximum A Posterior (MAP) problem as,

$$\hat{\mathbf{Y}} = \arg \max_{\mathbf{Y}} log P(\mathbf{X}|\mathbf{Y}) + log P(\mathbf{Y}) \quad (2)$$

where $\hat{\mathbf{Y}}$ is the predicted SR 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. Note that Equation 2 generally describes the optimization process that multiple SR $\hat{\mathbf{Y}}$ can be found. We can rewrite Equation 2 as Reference based SR,

$$\hat{\mathbf{Y_i}} = \arg \max_{\mathbf{Y}} log P(\mathbf{X}, \mathbf{R}_i|\mathbf{Y}) + log P(\mathbf{Y})$$
$$where \ \mathbf{R}_i \in \mathbb{S}_{\mathbf{R}} \ and \ \hat{\mathbf{Y_i}} \in \mathbb{S}_{\mathbf{Y}} \quad (3)$$

and we define the SR space as $\mathbb{S}_{\mathbf{Y}}$ and the reference space as $\mathbb{S}_{\mathbf{R}}$. The desired SR image can be randomly sampled from $\mathbb{S}_{\mathbf{R}}$ space given arbitrary reference image.

The complete architecture of the proposed *RefVAE* is shown in Figure 2, which includes the training (upper half of the figure) and testing (lower half of the figure) stages. Our proposed *RefVAE* takes arbitrary references $\mathbf{R}_i$ and LR images $\mathbf{X}$ for training and testing. It includes three components: 1) VGG Encoder, 2) CVAE (in the pink box), and 3) Image Decoder. The VGG Encoder is used to extract feature maps from LR and reference images. CVAE is used to transfer reference features to a hidden space for conditional feature sampling. In other words, CVAE generates a candidate feature pool where random feature vectors can be sampled as conditional features. The LR feature maps are split into mean and variance maps, then they compute with the conditional features to obtain the estimated features. Finally, the Image Decoder takes the estimated features for image reconstruction.

3.1. VGG Encoder

The VGG Encoder follows the structure of VGG-19 by keeping all convolution layers and discarding the fully connection layers. We directly use pre-trained VGG19 [34] to extract feature maps from references ($F_{\mathbf{R}} = G(\mathbf{R})$) and the bicubic upsampled LR images ($F_{\mathbf{X}} = G(\mathbf{X})$), where G stands for the process of VGG feature extraction. Inside the VGG Encoder, there are three maxpooling layers to down-sample the input image by 8×. Note that we resize arbitrary reference images to 256×256 before passing it through the
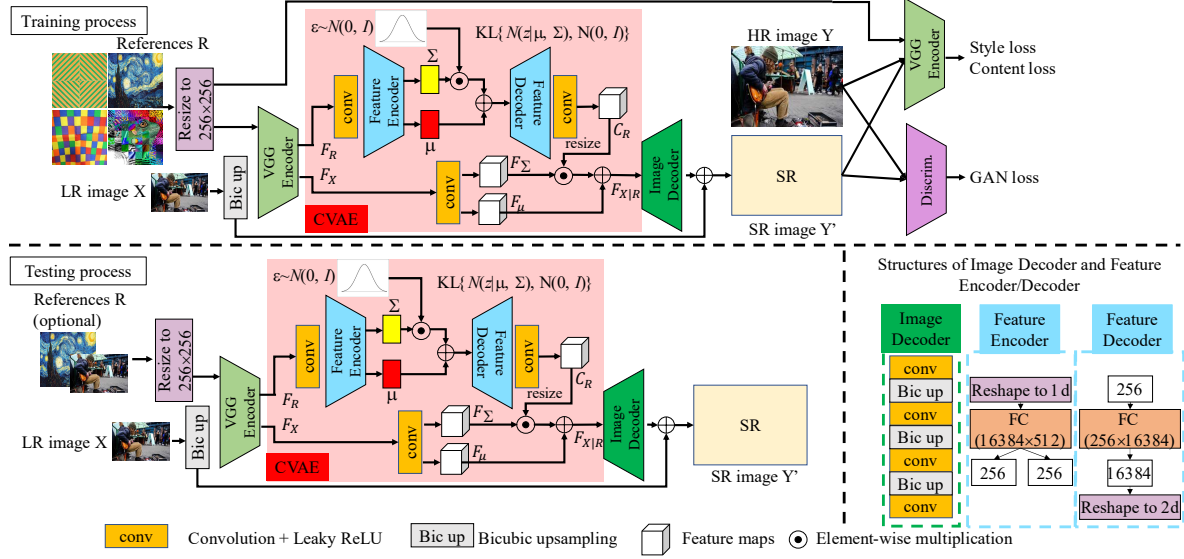
Figure 2: **The training and testing processes of the proposed RefVAE**. It consists of 1) VGG Encoder (the green box), 2) Conditional Variational Variational AutoEncoder (CVAE) (the pink box), 3) Image Decoder (the dark green box), and Discriminator (Discrim.) for loss computation. Detailed structures of the Image Decoder and Feature Encoder/Decoder are also shown on the right bottom corner.

VGG Encoder. Hence we have fixed reference feature maps as $F_{\mathbf{R}} \in \mathbb{R}^{32 \times 32 \times 512}$. For LR images, we initially upsample them by Bicubic interpolation to the desired size. The reason of using pre-trained VGG Encoder is that: 1) VGG was trained for general image classification that the extracted feature maps are generalized to images with different contents and 2) we want to project the LR and reference images to a same feature domain such that we can fuse their features together for super-resolution.

3.2. Conditional Variational AutoEncoder

We have the Conditional Variational AutoEncoder (CVAE) that projects the reference feature maps to a latent space to learn the hidden distribution via Feature Encoder. The Feature Decoder learns to transfer the reference features as conditions C_R for LR feature maps. The combination of Feature Encoder and Decoder (for detailed structures: see the right of Figure 2) forms the Variational Inference process. The idea of Variational Inference is to learn the generative model for the reference images that can be represented by a Gaussian model as $P(z|\mathbf{R}) = N \sim (z; \mu(\mathbf{R}), \Sigma(\mathbf{R}))$, where μ and $\Sigma = diag(\sigma_1^2, ..., \sigma_n^2)$ are the mean and variance of the learned Gaussian model. In other words, we learn the pixel inter-correlations and represent them as a probability model. We introduce randomness by the sampling process. As the Gaussian curve shown in Figure 2, we use a normal distribution $Q(z) = N \sim (0, I)$ to sample from the reference model as $z = \mu + \epsilon \cdot \sigma$. To ensure the learned probability model close to a normal distribution, we use KL divergence to optimize the model as,

$$D_{KL}(P(z|\mathbf{R})||Q(z)) = E[logP(z|\mathbf{R}) - Q(z)]$$
$$= \frac{1}{2}\left(-\sum_i(log\sigma_i^2 + 1) + \sum_i \sigma_i^2 + \sum_i \mu_i^2\right) \quad (4)$$

During the testing, we can discard the Feature Encoder as many existing VAE based image reconstruction [32, 16]. We can use vectors sampled from the normal distribution as conditional priors for super-resolution. We can also keep the Feature Encoder to allow user to define their own reference image for customized super-resolution. That is, we keep the Feature Encoder to extract specific prior distribution $P(z|\mathbf{R}) = N \sim (z; \mu(\mathbf{R}), \Sigma(\mathbf{R}))$ as prior conditions for super-resolution. Next, to obtain the conditional feature maps $C_{\mathbf{R}}$, we use one convolution block to project the learned distribution back to the spacial domain (we also resize it to the same size as the LR feature map via simple interpolation).

In order to transfer the conditional features to the LR feature map, we use a convolution block to learn the mean and variance (note that the mean and variance are the spatial statistics of the feature maps, rather than the variables of the Gaussian distribution) for the LR feature maps as F_μ and F_Σ. We then have the fused features as $F_{X|R} = C_R \cdot (1 + F_\Sigma) + F_\mu$.

3.3. Image Decoder

Finally, the Image Decoder learns to reconstruct the SR image from the fused feature maps $\hat{\mathbf{Y}}$. The structure of Image Decoder (dark green boxes in Figure 2) has a similar

structure as the VGG Encoder stacking three convolution layers followed by a simple bilinear interpolation.

3.4. Training losses

To train the proposed *RefVAE* to generate SR results with photo-realistic visual quality, we suggest to use a discriminator to reduce the perceptual distance between SR and ground truth images. We design the same discriminator as PatchGAN [38] and the adversarial loss is defined as,

$$L_{adv} = log[1 - D(\hat{\mathbf{Y}})] \qquad (5)$$

The idea of using style and content losses for style transfer [32, 16] is an efficient approach to transfer the reference style to the target image while preserving the content information. It is also suitable for reference based SR. We not only want the SR image close to the ground truth, but we also want it close to the reference image in terms of style similarity. In other words, we want to ensure the reference features to be transferred to the LR images.

Content loss For content loss, we extract features for the HR image $\mathbf{Y}$ and the SR $\hat{\mathbf{Y}}$ using VGG-19 [34] (we take the feature map at *relu4_1* layer). We refer these features to as $W_{\mathbf{Y}}$ and $W_{\hat{\mathbf{Y}}}$, respectively. For pixel-wised difference, we also have the L_1 loss between SR and HR and their down-sampled versions. Totally, we have the content loss as,

$$L_{\text{content}} = \left\|W_{\mathbf{Y}} - W_{\hat{\mathbf{Y}}}\right\|^1 + \left\|\mathbf{Y} - \hat{\mathbf{Y}}\right\|^1 + \left\|Down(\mathbf{Y}) - Down(\hat{\mathbf{Y}})\right\|^1 . \qquad (6)$$

where *Down* is the $\alpha\times$ bicubic down-sampling operator.

Style loss. To measure the style similarity, we use VGG-19 to extract feature maps (*relu1_2*, *relu2_2*, *relu3_4*, *relu4_1*) for the reference image $\mathbf{R}$ and the SR image $\hat{\mathbf{Y}}$ as $V_{\mathbf{R}}$ and $V_{\hat{\mathbf{Y}}}$. Similar to [32], we measure their mean and variance to align the SR features close to the reference features as:

$$L_{\text{style}} = \sum_i \left\|mean(V_{\mathbf{Y}}^i - V_{\hat{\mathbf{Y}}}^i)\right\|^1 + \left\|var(V_{\mathbf{Y}}^i - V_{\hat{\mathbf{Y}}}^i)\right\|^1 . \qquad (7)$$

where *mean* and *std* are the operations for calculating the mean and variance of the feature maps. In order to have SR images visually close to the HR image, the LPIPS loss [42] is used to measure perceptual differences as $L_{LPIPS}(\hat{\mathbf{Y}}, \mathbf{Y})$. The Total Variation loss $L_{TV}(\hat{\mathbf{Y}})$ is used to encourage smooth quality, for which we calculate the first-order horizontal and vertical pixel gradients as

$$L_{TV} = \sum_{i,j} \left((\hat{\mathbf{Y}}_{i,j-1} - \hat{\mathbf{Y}}_{i,j})^2 + (\hat{\mathbf{Y}}_{i-1,j} - \hat{\mathbf{Y}}_{i,j})^2\right)^\beta . \qquad (8)$$

Total loss. We have the total loss as,

$$L_{\text{style}} = \lambda_{content} L_{\text{content}} + \lambda_{style} L_{\text{style}} + \lambda_{LPIPS} L_{\text{LPIPS}} \\ + \lambda_{TV} L_{\text{TV}} + \lambda_{KL} L_{\text{KL}} . \qquad (9)$$
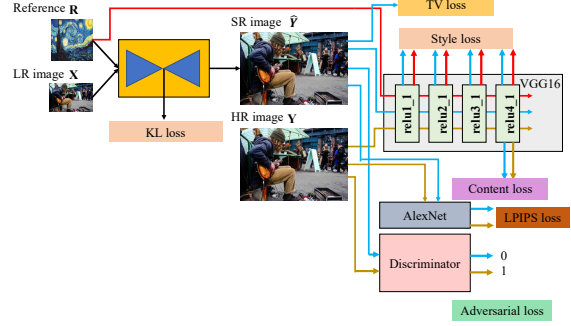

Figure 3: **Training Losses in the proposed *RefVAE***. It consists of 1) Content loss, 2) Style loss, 3) KL loss, 4) TV loss, 5) LPIPS loss and 6) Adversarial loss.

where $\lambda_{content}$, λ_{style}, λ_{LPIPS}, λ_{TV} and λ_{KL} are the weighting parameters for content loss, style loss, LPIPS loss, TV loss and KL loss. Figure 3 shows a summary of all the loss terms for readers' reference.

4. Experiments

4.1. Data Preparation and Network Details

We trained our model on DIV2K [37] and Flickr2K [21] datasets. They both contain images with resolution larger than 1000×1000. We extracted LR and HR patches from the training dataset with the size of 32×32 and $32\alpha \times 32\alpha$, respectively, where α=8 is the upsampling factor. The down-sampling process is done by bicubic interpolation. For the reference image, we used Wikiart [1] that is widely used in style transfer. The testing datasets included Set5 [5], Set14 [41], Urban100 [14] and DIV2K validation [37].

We conducted our experiments using Pytorch 1.7 on a PC with one NVIDIA GTX1080Ti GPU. During the training, we set the learning rate to 0.0001 for all layers. The batch size was set to 16 and we trained it for 5×10^4 iterations. For optimization, we used Adam with the momentum equal to 0.9 and the weight decay of 0.0001. During the training, we set $\lambda_{content} = 1$, $\lambda_{style} = 10$, $\lambda_{LPIPS} = 1$, $\lambda_{TV} = 1$ and $\lambda_{KL} = 1$, which are the weighting parameters for content loss, style loss, LPIPS loss, TV loss and KL loss. The executive codes and more experimental results can be found in: https://github.com/Holmes-Alan/RefVAE

Metrics. For evaluation, we have used several metrics: **PSNR** measures the average pixel differences between ground truth and estimation. **SSIM** [40] measures the structural similarity between ground truth and estimation. [1] **LPIPS** [42] measures the perceptual similarity between ground truth and estimation. **Diverse score** measures the

---

[1]To calculate PSNR and SSIM, we first convert the SR and ground truth images from RGB to YCbCr and take the Y channel for computation.
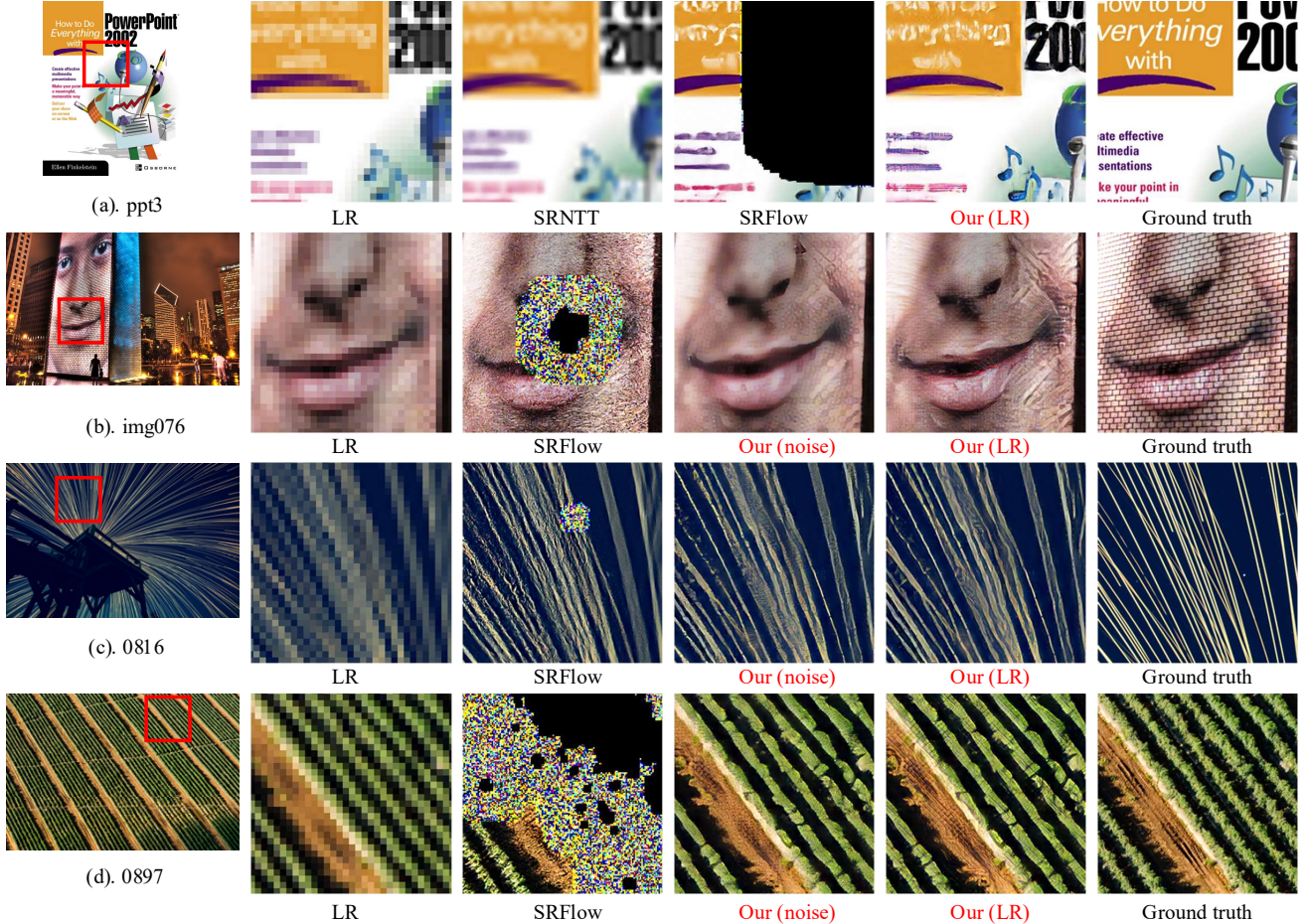
Figure 4: **Visualization of 8× SR on different datasets using different SR methods**. (a) is from Set14, (b) is from Urban100, (c) and (d) are from DIV2K validation.

Table 1: **Comparison with state-of-the-art methods**. We compare with SRNTT and SRFlow on 8× super-resolution on PSNR (dB), SSIM and LPIPS.

Methods	Set5			Set14			BSD100			Urban100			DIV2K validation		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Bicubic	24.39	0.657	0.537	23.19	0.568	0.630	23.67	0.547	0.713	21.24	0.516	0.686	25.17	0.664	0.584
SRNTT(HR)	24.69	0.674	0.492	23.16	0.564	0.613	23.63	0.542	0.687	-	-	-	-	-	-
SRNTT(LR)	24.36	0.655	0.522	23.16	0.564	0.613	23.65	0.543	0.691	-	-	-	-	-	-
SRFlow	24.18	0.650	0.237	22.13	0.513	0.331	22.96	0.499	0.425	20.676	0.531	0.300	24.53	0.616	0.272
Our(LR)	25.87	0.723	**0.214**	24.15	0.598	**0.323**	24.15	0.558	**0.376**	21.74	0.575	0.334	24.75	0.641	0.232
Our(HR)	25.92	**0.724**	0.218	24.20	**0.601**	0.327	24.18	**0.561**	0.386	21.80	0.575	**0.332**	24.94	**0.650**	**0.201**
Our(random)	**25.94**	0.721	0.224	**24.28**	**0.601**	0.328	**24.25**	**0.561**	0.384	**21.85**	**0.576**	0.358	**25.00**	**0.650**	0.202

spanning of the SR space. We follow the same measurement defined by NTIRE2021 SR space challenge [29]. We sampled 10 images, and densely calculated a metric between the samples and the ground truth. To obtain the local best we pixel-wisely selected the best score out of the 10 samples and took the full image's average. The global best was obtained by averaging the whole image's score and selecting the best. Finally, we calculated the score as: (global best - local best)/(global best) × 100.

4.2. Compare with state-of-the-art methods

Our proposed *RefVAE* is for diverse image super-resolution that is able to generate many SR results. We consider two types of related SR methods for comparison: reference based image SR and generative model based image SR. The SR approaches we compare with are: **SRNTT** [45], reference based SR using similar image as reference for on-line warping, and **SRFlow** [30], normalizing flow based SR
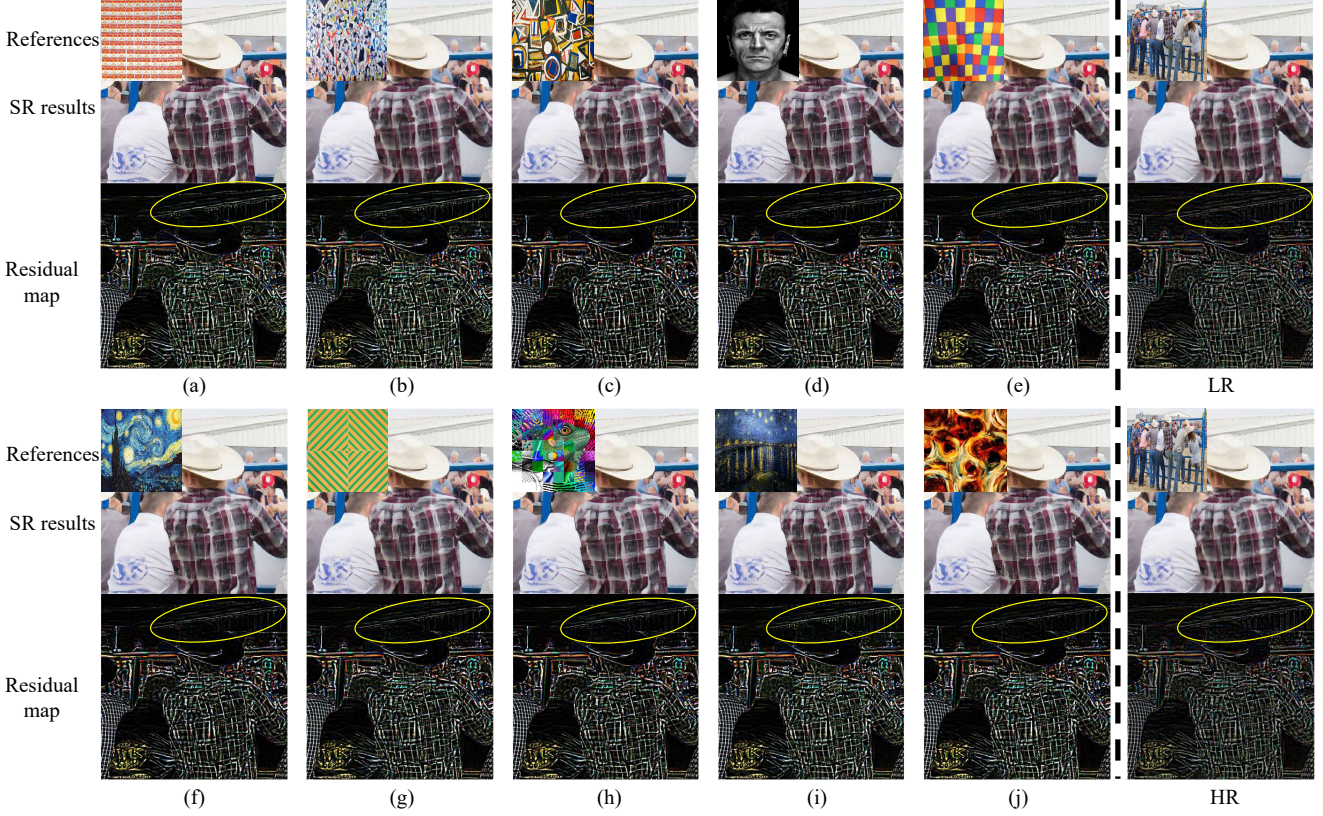
Figure 5: **Visualization of 8× SR on using different reference images**. To visualize the differences, we subtract the SR to the ground truth HR image to obtain the residual map. Then we multiple the residual map by 5 to enhance the differences. We also show two examples of using LR and HR images as references for SR. The significant differences are marked in yellow circles.

that predicts multiple SR results. [2]

In Table 1, we show the overall comparison with SRNTT and SRFlow on 8× image super-resolution on Set5, Set14, BSD100, Urban100 and DIV2K validation datasets. It is seen that our proposed *RefVAE* can achieve the best PSNR, SSIM and LPIPS by about 2 dB, 0.1 and 0.01, respectively. Note that SRNTT [45] was originally trained and tested on specific CUFED5 dataset [45] for which each image has five images with similar contents. For a fair comparison, we tested SRNTT by using LR and HR images as reference for SR, which are SRNTT(LR) and SRNTT(HR). [3] Similarly, we also tested our proposed *RefVAE* by using LR and HR images as Our(LR) and Our(HR), respectively. Visually, we show four examples in Figure 4. It can be seen that our proposed method can restore the details better than others. Especially, it is seen that SRFlow collapses on several cases.

For example, SRFlow generated a large black region on *ppt3*. It also produces holes on *img076*, *0816* and *0897* with noises. For SRNTT, it generated blurry results without restoring sharp edges. On the other hand, our proposed *RefVAE* can generate plausible SR results with sharp edges and textures.

More importantly, we are interested in the ability of expansion of SR space. Since our proposed *RefVAE* can take arbitrary references for super-resolution, we can measure the diversity by using **Diverse score** defined in NTIRE2021 SR space challenge [29].

As discussed in Section III, our proposed *RefVAE* can take random noise or external images as references. In Table 2, we randomly chose 10 images from WikiArt as references for SR. We also use LR image itself and HR image as references. We list the results on SR estimation as *SR PSNR*, *SR SSIM* and *SR LPIPS*. It can be seen that the proposed *RefVAE* can generate different SR images (different PSNR and SSIM) given different references. It demonstrates that the proposed *RefVAE* can expand the SR space by a large margin. In the meantime, we do not want to distort SR images to be far away from the ground truth LR images. Hence, we down-

---

[2]Note that since most of state-of-the-art SR approaches are dedicated for producing one prediction with high PSNR and do not consider the possibility of multiple SR results, it is not our intention to compete with those approaches in terms of PSNR and SSIM.

[3]Note also that SRNTT requires a lot of memory for online warping. Larger image SR cannot be done on one GPU, hence we did not have the results for Urban100 and DIV2K.

Table 2: **Analysis on SR space** of *RefVAE* on DIV2K dataset. The PSNR, SSIM and LPIPS on SR estimation are indicated as *SR PSNR*, *SR SSIM* and *SR LPIPS*, while for LR estimation they are respectively indicated as *LR PSNR* and *LR SSIM*. The overall estimation using SRFlow and our methods are shown in the 2nd half of the table.

Reference	SR PSNR ↑	SR SSIM ↑	SR LPIPS ↓	LR PSNR ↑	LR SSIM ↑
Ref. 1	24.27	0.616	0.310	42.90	0.996
Ref. 2	25.18	0.661	0.306	44.55	0.997
Ref. 3	25.03	0.654	0.304	44.37	0.997
Ref. 4	24.76	0.644	0.302	43.93	0.997
Ref. 5	25.28	0.664	0.314	44.80	0.997
Ref. 6	25.17	0.662	0.306	44.81	0.997
Ref. 7	24.62	0.627	0.298	43.43	0.996
Ref. 8	25.04	0.659	0.307	44.26	0.997
Ref. 9	24.27	0.611	0.320	42.86	0.996
Ref. 10	24.42	0.615	0.310	42.30	0.996
LR	25.31	0.664	0.306	46.21	0.997
HR	25.40	0.665	0.308	45.63	0.997
Overall Estimation					
SRFlow Div. score: 10.07	24.46	0.615	0.328	**51.18**	0.996
Our (Reference) Div. score: 14.80	24.89	0.641	**0.308**	44.17	**0.997**
Our (Random) Div. score: **14.91**	**24.91**	**0.642**	0.321	44.77	**0.997**

Table 3: **Ablation study** of *RefVAE* on DIV2K dataset on 8× super-resolution. We report the average across datasets.

Model	Components			Eval.			
	CVAE	SC	Dicri.	PSNR ↑	SSIM ↑	LPIPS ↓	Div. ↑
Our	-	-	-	**25.23**	**0.650**	0.412	0
	✓	-	-	24.82	0.641	0.338	12.326
	✓	✓	-	24.84	0.643	0.400	14.142
Final	✓	✓	✓	24.75	0.640	**0.272**	**14.992**

sample SR images by bicubic to estimate PSNR and SSIM on the LR space as *LR PSNR* and *LR SSIM*, respectively. It can be seen that different SR estimation can well preserve LR information with PSNR over 40 dB.

Overall, we have calculated the **Diverse score** of SR-Flow and our approach in Table 2. Given different references, we have two results as entitled Our(Reference) and Our(Random). It can be seen that our approach can outperform SRFlow in terms of PSNR, SSIM and **Diverse score**. We also show SR results of using 12 different references in Figure 5. To visualize the differences among multiple SR images, we subtracted the SR by the ground truth HR images to obtain the residual maps, then we multiplied the residual values by 5 to show the differences. We can observe the differences around the edge and texture regions from the residual maps (marked in yellow circles), e.g., the patterns of the T-shirt and the roof of the building. In the meantime, we also visualize the SR results using LR and HR images. It can be seen that using LR or HR as references can better reduce the residues.

4.3. Ablation study

Let us evaluate various key components of the *RefVAE*: (1) the Conditional Variational AutoEncoder (CVAE), (2)

Figure 6: **Weak cases of using proposed *RefVAE*.** We show two 8× SR examples from BSD100 dataset. It can be found that the facial details cannot be preserved.

the use of style and content losses (SC loss), (3) the use of discriminator (Discri.), To examine their impact, we started with a simple network structure consisting Image Encoder and Decoder only, and we progressively added all components. Table 3 reports the results of 8× SR on DIV2K validation datasets. We have included PSNR, SSIM, LPIPS and *Div.* (**Diverse score**) in the table. From the values of *Div.*, we can find that using CVAE module can introduce randomness into the network to generate multiple SR results. When the style and content losses were used for training, we can further maximize the **Diverse score** to expand the SR space. From the values of PSNR, SSIM and LPIPS, using CVAE, SC loss and Discri can reduce the PSNR and SSIM, but they can improve the LPIPS value, which indicates better quality.

4.4. Weak cases

In terms of perceptual quality, our proposed method does not perform too well on images with smaller size. We show two examples in Figure 6. It can be seen that the fine facial expression cannot be ideally restored by our approach. To resolve this problem, we could further study in depth of the latent representation of the Variational AUtoEncoder. One other solution is to use hierarchical latent space manipulation. This is really a fruitful direction for research.

5. Conclusion

In this paper, we introduce a novel approach for reference based image super-resolution *RefVAE*. Unlike other reference based SR approaches, our proposed method can take any image with arbitrary content or even random noise as reference for image super-resolution. It is also unlike other generative model based SR approaches, *RefVAE* can expand the SR space, so that multiple unique SR images can be generated. Furthermore, *RefVAE* does not suffer from the model collapse or generate bizarre patterns like other generative model based approaches (examples are shown in Figure 4). Our analysis shows that *RefVAE* can better preserve the LR information and generate SR image with better perceptual quality. Possible future work includes the super-resolution space exploration by probing the latent vector or by image quantization.

References

[1] K. nichol. painter by numbers, wikiart. https://www.kaggle.com/c/painter-by-numbers, 2016. 5

[2] Nvidia dlss 2.0: A big leap in ai rendering. https://www.nvidia.com/en-gb/geforce/news/nvidia-dlss-2-0-a-big-leap-in-ai-rendering/, March, 2020. 1

[3] Amd fidelityfx. https://www.amd.com/en/technologies/radeon-software-fidelityfx, October, 2020. 1

[4] Yuval Bahat and Tomer Michaeli. Explorable super resolution. In *CVPR*, pages 2713–2722. IEEE, 2020. 2

[5] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pages 135.1–135.10. BMVA Press, 2012. 5

[6] Y. Blau and T. Michaeli. The perception-distortion trade-off. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 1

[7] Marcel C. Bühler, Andrés Romero, and Radu Timofte. Deepsee: Deep disentangled semantic explorative extreme super-resolution. In *ACCV*, volume 12625 of *Lecture Notes in Computer Science*, pages 624–642. Springer, 2020. 2

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 2

[9] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, Feb 2016. 2

[10] Jesse Engel, Matthew Hoffman, and Adam Roberts. Latent constraints: Learning to generate conditionally from unconditional generative models. *CoRR*, abs/1711.05772, 2017. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 2

[12] Lei Han Ziwei Xu Haoqian Wang Yebin Liu Haitian Zheng, Mengqi Ji and Lu Fang. Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 138.1–138.13. BMVA Press, September 2017. 1, 3

[13] M. Haris, G. Shakhnarovich, and N. Ukita. Deep back-projection networks for single image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[14] J. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, June 2015. 5

[15] Kwok-Wai Hung and Wan-Chi Siu. Robust soft-decision interpolation using weighted least squares. *IEEE Transactions on Image Processing*, 21(3):1061–1069, 2012. 1, 3

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Eur. Conf. Comput. Vis.*, 2016. 4, 5

[17] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 2

[18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. *CoRR*, abs/1511.04587, 2015. 2

[19] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. *CoRR*, abs/1704.03915, 2017. 2

[20] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. 1, 2

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017. 1, 2, 5

[22] Zhi-Song Liu and Wan-Chi Siu. Cascaded random forests for fast image super-resolution. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2531–2535, Oct 2018. 2, 3

[23] Zhi-Song Liu, Wan-Chi Siu, and Yui-Lam Chan. Fast image super-resolution via randomized multi-split forests. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4, May 2017. 2

[24] Zhi-Song Liu, Wan-Chi Siu, and Yui-Lam Chan. Reference based face super-resolution. *IEEE Access*, 7:129112–129126, 2019. 2, 3

[25] Zhi-Song Liu, Wan-Chi Siu, and Yui-Lam Chan. Photo-realistic image super-resolution via variational autoencoders. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020. 2

[26] Zhi-Song Liu, Wan-Chi Siu, Wang Li-Wen, Li Chu-Tak, Marie-Paule Cani, and Yui-Lam Chan. Unsupervised real image super-resolution via generative variational autoencoder. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshop(CVPRW)*, June 2020. 2

[27] Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, and Wan-Chi Siu. Hierarchical back projection network for image super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2041–2050, 2019. 2

[28] Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, Wan-Chi Siu, and Yui-Lam Chan. Image super-resolution via attention based back projection networks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3517–3525, 2019. 2

[29] Andreas Lugmayr, Martin Danelljan, and et al. Radu Timofte. Ntire 2021 learning the super-resolution space challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 3, 6, 7

[30] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *ECCV*, pages 715–732. Springer, 2020. 2, 6

[31] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 191–207, Cham, 2020. Springer International Publishing. 2

[32] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 4, 5

[33] N. C. Rakotonirina and A. Rasoanaivo. Esrgan+ : Further improving enhanced super-resolution generative adversarial network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3637–3641, 2020. 1, 2

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. Learn. Represent.*, 2015. 3, 5

[35] Y. Tan, H. Zheng, Y. Zhu, X. Yuan, X. Lin, D. Brady, and L. Fang. Crossnet++: Cross-scale large-parallax warping for reference-based super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 3

[36] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. volume 9006, pages 111–126, 04 2015. 2

[37] Radu Timofte and et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2017*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pages 1110–1121. IEEE Computer Society, 8 2017. 5

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5

[39] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. ESRGAN: enhanced super-resolution generative adversarial networks. *CoRR*, abs/1809.00219, 2018. 1, 2

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. 5

[41] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. volume 6920, pages 711–730, 06 2010. 5

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[43] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. *CoRR*, abs/1807.02758, 2018. 2

[44] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. *CoRR*, abs/1802.08797, 2018. 2

[45] Z. Zhang, Z. Wang, Z. Lin, and H. Qi. Image super-resolution by neural texture transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7974–7983, 2019. 1, 3, 6, 7

[46] Haitian Zheng, Mengqi Ji, Haoqian Wang, Yebin Liu, and Lu Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 88–104, 2018. 3