pSp: A StyleGAN Encoder for Super Resolution

Xinyi Shang, Pinxin Qian, Jun Liu, Zhongxuan Sun, JunHao Zhao

[1] Fujian Key Laboratory of Sensing and Computing for Smart City,
School of Informatics, Xiamen University,
Xiamen, China

Abstract

The primary aim of single-image super-resolution is to construct a high-resolution (HR) image from a corresponding low-resolution (LR) input. Previous approaches have achieved very high-quality results, but they are unable to accurately retain identity on face images. We present a image-to-image translation framework addressing this problem, pSp(Pixel2Style2Pixel). Firstly, pSp framework is based on a novel encoder network that directly generates a series of style feature vectors that can reconstruct a given image while preserving the identity and other attributes, forming the extended $\mathcal{W}+$ latent space. Then ,$\mathcal{W}+$ are fed into a pretrained StyleGAN generator. Our encoder can directly embed real images into $\mathcal{W}+$, with no additional optimization. Notably, the advantage of the intermediate style representation is the inherent support of multi-modal synthesis for tasks such as low-resolution images. Moreover, We introduce a dedicated identity loss which is shown to achieve improved performance in the reconstruction of an input image. We demonstrate that pSp can be trained to construct high-resolution images from corresponding low-resolution images. Finally, We show experimental results demonstrating the efficacy of our approach in the domain of face super resolution.

Keywords super-resolution, Generative adversarial network, Identity preserving

Introduction

In many areas (such as medicine, astronomy, microscopy, and satellite imagery), sharp, high-resolution images are difficult to obtain due to issues of cost, hardware restriction, or memory limitations (Singh and Singh 2016). This leads to the capture of blurry, low-resolution images instead. In other cases, images could be old and therefore blurry, or even in a modern context, an image could be out of focus or a person could be in the background. In addition to being visually unappealing, this impairs the use of downstream analysis methods (such as image segmentation, action recognition, or disease diagnosis) which depend on having high-resolution images (Ronneberger, Fischer, and Brox 2015; Simonyan and Zisserman 2014). In addition, as consumer laptop, phone, and television screen resolution has increased over recent years, popular demand for sharp images and video has surged. This has motivated recent interest in the computer vision task of image super-resolution, the creation of realistic high-resolution (henceforth HR) images that a given low-resolution (LR) input image could correspond to. In this work, we aim to transform blurry, low-resolution images into sharp, realistic, high-resolution images. Notably, the output image can retain identity. Here, we focus on images of faces, but our technique is generally applicable.

Several methods have been proposed to improve the visual quality of SR results. For instance, perceptual loss (Zhang et al. 2018b) is proposed to optimize super-resolution model in a feature space instead of pixel space. Generative adversarial network is introduced to SR by (Ledig et al. 2016) to encourage the network to favor solutions that look more like natural images. StyleGAN (Karras, Laine, and Aila 2020) proposes a novel stylebased generator architecture and attains state-of-the-art visual quality on high-resolution images. Moreover, it has been demonstrated that it has a disentangled latent space, $\mathcal{W}$ (Yang, Shen, and Zhou 2019), obtained from the initial latent space Z via a Multi-Layer Perceptron (MLP) mapping network, which may offer control and editing capabilities. pix2pixHD (Wang et al. 2018) is able to obtain good results. However, visually, its results appear less photo-realistic. Although PULSE (Menon et al. 2020) is able to achieve very high-quality results, they are unable to accurately retain identity.

In this paper, we focus on latent space embedding, which aims at the retrieval of the latent vector that generates a desired image. Firstly, We do so by introducing a novel encoder architecture tasked with encoding an arbitrary image directly into $\mathcal{W}+$, which can reconstruct a given image while preserving the identity and other attributes. The encoder is based on a Feature Pyramid Network (Lin et al. 2017), where style feature vectors are extracted from different pyramid scales and inserted directly into a fixed, pretrained StyleGAN generator in correspondence to their spatial scales. Besides the simplification of the training process, as no adversary discriminator needs to be trained, using a pretrained StyleGAN generator offers several intriguing advantages over previous works. Our encoder into $\mathcal{W}+$, together with the StyleGAN decoder, form a encoder-decoder network that benefits many image-to-image translation tasks. Then, We introduce a identity loss which is shown to achieve improved performance in the reconstruc-

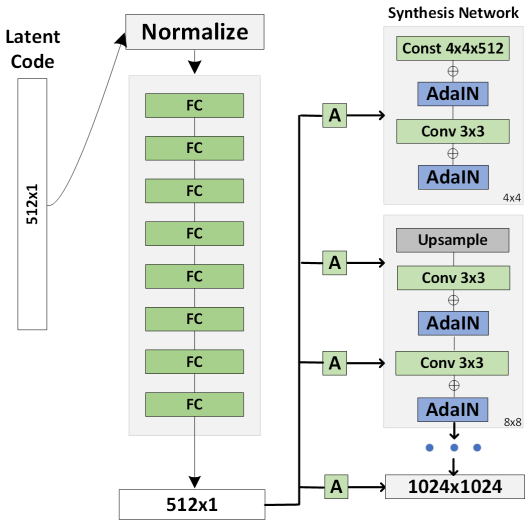

Figure 1: The mapping network f consists of 8 layers and the synthesis network g consists of 18 layers—two for each resolution ($4^2 - 1024^2$). The output of the last layer is converted to RGB using a separate 1×1 convolution, similar to Karraset al.

tion of an input image. Focusing on face images, we demonstrateour method's ability to successfully reconstruct a given image while preserving the identity and other attributes and retain identity on face images. In a sense, our method performs Pixel2Style2Pixel translation, as every image is first encoded into style feature vectors and then into an image, and is therefore dubbed pSp.

Related Work

Latent Space Embedding With the rapid evolution of GANs, many works have tried to understand and control their latent space. A specific task that has received substantial attention is GAN Inversion — where the latent vector from which a pretrained GAN most accurately reconstructs a given, known image, is sought. Motivated by its state-of-the-art image quality and latent space semantic richness, many recent works have used StyleGAN for this task (Karras, Laine, and Aila 2020). Generally, inversion methods either directly optimize the latent vector to minimize the error for the given image (Abdal, Qin, and Wonka 2020), train an encoder to map the given image to the latent space (Antonia et al. 2018), or use a hybrid approach combining both (Zhu et al. 2020). Typically, methods performing optimization are superior in reconstruction quality to a learned encoder mapping, but are costly and require a substantially longer time.

Focusing on the more general task of latent space embedding, Nitzan et al. (Nitzan et al. 2020) trained an encoder to infer a latent vector from which StyleGAN can directly generate an image with the identity of one image and the pose, expression, and illumination of another. While this shows the potential of latent embedding, their method solves only a specific application and cannot be used to solve other image-to-image translation tasks.

Image-to-Image Translation techniques aim at learning a conditional image generation function that maps an input image of a source domain to a corresponding image of a target domain. Isola et al. (Isola et al. 2017) first introduced the use of conditional GANs to solve various image-toimage translation tasks. Since then, their work has been extended for many scenarios: high-resolution synthesis (Wang et al. 2018), unsupervised learning (Liu, Breuel, and Kautz 2017; Lira et al. 2020), multi-modal image synthesis (Huang et al. 2018; Choi et al. 2020), multi-domain image synthesis (Choi et al. 2018, 2020), and conditional image synthesis (Yang, Shen, and Zhou 2019; Zhu et al. 2019; Chen et al. 2020; Liu et al. 2019). The aforementioned works have constructed dedicated architectures for their tasks which require training the generator network. This is in contrast to our method that utilizes a fixed pretrained StyleGAN generator, enjoying its state-of-the-art image quality.

Super-Resolution Recently, supervised neural networks have come to dominate current work in super-resolution. Dong et al. (Dong et al. 2014) proposed the first CNN architecture to learn this non-linear LR to HR mapping using pairs of HR-LR images. Several groups have attempted to improve the upsampling step by utilizing sub-pixel convolutions and transposed convolutions (Shi et al. 2016). Furthermore, the application of ResNet architectures to super-resolution (started by SRResNet (Ledig et al. 2016)), has yielded substantial improvement over more traditional convolutional neural network architectures. In particular, the use of residual structures allowed for the training of larger networks.

Method

StyleGAN Generator

The complete architecture is illustrated in Figure1. Given a latent code $\mathbf{z}$ in the input latent space $\mathcal{Z}$, a non-linear mapping network $f : \mathcal{Z} \rightarrow \mathcal{W}$ first produces $w \in \mathcal{W}$. For simplicity, we set the dimensionality of both spaces to 512, and the mapping f is implemented using an 8-layer MLP. Learned affine transformations then specialize $\mathbf{w}$ to $\mathbf{y} = (\mathbf{y}_s, \mathbf{y}_b)$ that control adaptive instance normalization (AdaIN) (Isola et al. 2017; Lin et al. 2017; Huang et al. 2020; Karras et al. 2017) operations after each convolution layer of the synthesis network g-shown in Figure 3.The AdaIN operation in is defined as

$$\mathrm{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i}$$

where each feature map $\mathbf{x}_i$ is normalized separately, and then scaled and biased using the corresponding scalar components from style $\mathbf{y}$. Thus the dimensionality of $\mathbf{y}$ is twice the number of feature maps on that layer.

The pSp FrameWork

Our pSp framework builds upon the representative power of a pretrained StyleGAN generator and the $\mathcal{W}+$ latent space. To utilize this feature representation one needs a strong encoder that is able to match each input image to an accurate encoding in the latent domain. In StyleGAN, the authors have shown that the different style inputs correspond

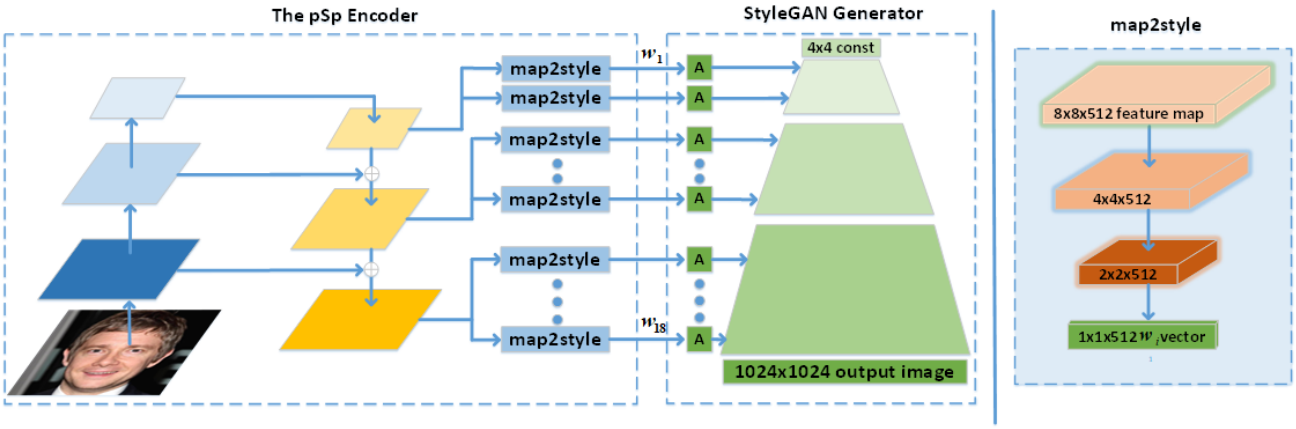

Figure 2: The architecture of our model.

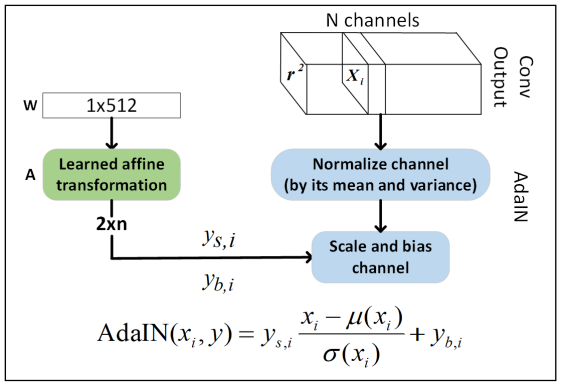

Figure 3: AdaIN

to different levels of detail, which are roughly divided into three groups — coarse, medium, and fine. Following this observation, in pSp, we extend an encoder backbone with a feature pyramid (Lin et al. 2017), generating three levels of feature maps from which styles are extracted using a simple intermediate network—map2style – shown in Figure 2. The styles, aligned with the hierarchical representation, are then fed into the generator in correspondence to their scale to generate the output image, thus completing the translation from input pixels to output pixels, through the intermediate style feature representation. Therefore, our architecture, pSp, is an end-to-end image-to-image translation framework. The complete architecture is illustrated in Figure2. We note that while we found the feature pyramid to best match the StyleGAN architecture, other possible variations could also work. For example, generating all the style vectors from the largest feature map would mostly affect the model size without hindering model accuracy. Conversely, generating the style vectors from the smallest feature map is also feasible without limiting performance as long as its dimensionality is large enough.

Loss Functions

Our encoder is trained using a weighted combination of several objectives. First, we utilize the pixel-wise $\mathcal{L}_2$ loss,

$$\mathcal{L}_2(\mathbf{x}) = \|\mathbf{x} - pSp(\mathbf{x})\|_2$$

where $\mathbf{x}$ denotes the input image and $pSp(\mathbf{x}) = G(E(\mathbf{x}))$ is the output returned by pSp defined by the encoder network, $E(\cdot)$, and generator network, $G(\cdot)$. In addition, to learn perceptual similarities, we utilize the LPIPS (Zhang et al. 2018a) loss, which has been shown to better preserve image quality compared to the more standard perceptual loss. Formally,

$$\mathcal{L}_{\text{LPIPS}}(\mathbf{x}) = \|F(\mathbf{x}) - F(pSp(\mathbf{x}))\|_2$$

where $F(\cdot)$ denotes the perceptual feature extractor.

The Identity Loss

One of the main challenges of face generation tasks is the ability to preserve identity between the input and output images. Since identity preservation is a crucial part of face reconstruction tasks, it is important to integrate this objective into the overall loss function. Therefore, as the aforementioned loss functions are less sensitive to the preservation of facial identity, we incorporate a dedicated recognition loss measuring the cosine similarity between the output image and its source,

$$\mathcal{L}_{\text{ID}}(\mathbf{x}) = 1 - < R(\mathbf{x}), R(pSp(\mathbf{x})) >$$

where R is a pretrained ArcFace (Deng, Guo, and Zafeiriou 2018) network for face recognition. The input image, $\mathbf{x}$, and corresponding generated image, $pSp(\mathbf{x})$, are cropped around the face and resized to 112 × 112 before being fed into R.

In summary, the total loss function is defined as

$$\mathcal{L}(\mathbf{x}) = \lambda_1\mathcal{L}_2(\mathbf{x}) + \lambda_2\mathcal{L}_{\text{LPIPS}}(\mathbf{x}) + \lambda_3\mathcal{L}_{\text{ID}}(\mathbf{x})$$

where λ_1, λ_2, and λ_3 are constants defining the loss weights.

Experiment

Datasets

We evaluated our procedure on the well-known highresolution face dataset CelebA HQ (Karras et al. 2017), which contains 30,000 high quality images. (Note: this is not to be confused with CelebA, which is of substantially lower resolution.) We use a standard train-test split of the dataset, resulting in approximately 24,000 training images.

Baselines

Here we show that our framework can be used to construct high-resolution (HR) facial images from corresponding low-resolution (LR) input images. PULSE approaches this task in an unsupervised manner. Specifically, for a given LR input image, PULSE traverses the HR image manifold in search of an image that downscales to the original LR image. Although PULSE takes an unsupervised approach to this problem, in this work we focus on applying pSp in a supervised manner for solving this task as obtaining paired data is immediate. We show that our method achieves comparable results, especially with respect to identity preservation.

Methodology and details

We train our super-resolution model in a supervised fashion, where for each input, we perform random bicubic down-sampling of $\times 1$ (i.e. no sub-sampling), $\times 2$; $\times 4$; $\times 8$ and $\times 16$ and set the original, full resolution image as the target.

Qualitative Image Results

Figure 4 demonstrates the visual quality of the resulting images from our method along with those of the previous approaches. Although PULSE is able to achieve very high-quality results due to their usage of StyleGAN to generate images, they are unable to accurately retain identity even when performing down-sampling of $\times 8$. Contrary to these previous works, we are able to obtain high-quality, photo-realistic images while successfully preserving identity, even when down-sampling by $\times 32$.

Quantitative Comparison

Table 1: MOS Score for various algorithms at 128 × 128. Higher is better.

Methods	**FSRNet**	**PULSE**	**pix2pixHD**	**pSp**
MOS	2.92	3.60	3.67	3.70

Here we present a quantitative comparison with state-ofthe-art face super-resolution methods. We conducted a mean-opinion-score (MOS) test as is common in the perceptual super-resolution literature (Duan et al. 2020). For this, we had 40 raters examine images upscaled by 5 different methods (FSRNet, PULSE, pix2pixHD and our pSp). For this comparison, we used a scale factor of 8 and a maximum resolution of 128×128, despite our method's ability to go substantially higher, due to this being the maximum limit for the competing methods. After being exposed to 20 examples

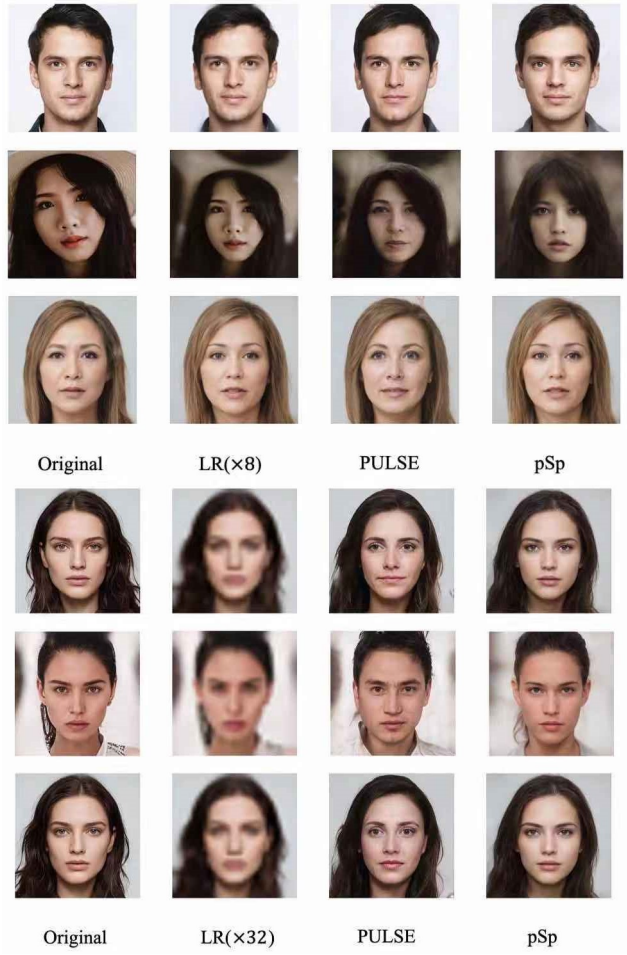

Figure 4: Visual comparison of super resolution results on CelebA-HQ.

of a 1 (worst) rating exemplified by nearest-neighbors up-sampling, and a 5 (best) rating exemplified by high-quality HR images, raters provided a score from 1-5 for each of the 240 images. All images fell within the appropriate $\epsilon = 1e{-}3$ for the downscaling loss. The results are displayed in Table 1.

pSp outperformed the other methods and its score approached that of the HR dataset. Note that the HR's 3.74 average image quality reflects the fact that some of the HR images in the dataset had noticeable artifacts. All pairwise differences were highly statistically significant by the Mann-Whitney-U test. The results demonstrate that pSp outperforms current methods in generating perceptually convincing images that downscale correctly.

Conclusion

In this work, we proposed a novel encoder architecture that can be used to directly map a face image into the $\mathcal{W}+$ latent space with no optimization required. The encoder architecture, motivated by StyleGAN, consists of a hierarchy of three levels that correspond to the coarse, medium, and fine

groupings of the 18 style vectors defining the input in the $\mathcal{W}+$ latent space. Styles are then extracted from the encoder in a hierarchical fashion and fed into the corresponding inputs of a fixed StyleGAN generator. Notably, our network is trained with an identity similarity loss, which encourages better preservation of identity compared to previous direct approaches. Finally, We show experimental results demonstrating the efficacy of our approach in the domain of face super resolution and our method can retain identity on face images.

References

Abdal, R.; Qin, Y.; and Wonka, P. 2020. Image2StyleGAN++: How to Edit the Embedded Images? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Antonia; Creswell; Anil; Anthony; and Bharath. 2018. Inverting the Generator of a Generative Adversarial Network. *IEEE transactions on neural networks and learning systems* .

Chen, S.; Su, W.; Gao, L.; Xia, S.; and Fu, H. 2020. DeepFaceDrawing. *ACM Transactions on Graphics (TOG)* .

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8188–8197.

Deng, J.; Guo, J.; and Zafeiriou, S. 2018. ArcFace: Additive Angular Margin Loss for Deep Face Recognition .

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a Deep Convolutional Network for Image Super-Resolution .

Duan, Y.; Liu, Y.; Wang, R.; Yao, D.; and Zhang, H. 2020. Progressive face super-resolution via learning prior information. *Journal of Physics: Conference Series* 1651(1): 012127 (6pp).

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 172–189.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5901–5910.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* .

Karras, T.; Laine, S.; and Aila, T. 2020. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(99): 1–1.

Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; and Wang, Z. a. 2016. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network .

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lira, W.; Merz, J.; Ritchie, D.; Cohen-Or, D.; and Zhang, H. 2020. GANHopper: Multi-Hop GAN for Unsupervised Image-to-Image Translation .

Liu, M. Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised Image-to-Image Translation Networks .

Liu, X.; Yin, G.; Shao, J.; Wang, X.; and Li, H. 2019. Learning to Predict Layout-to-image Conditional Convolutions for Semantic Image Synthesis .

Menon, S.; Damian, A.; Hu, S.; Ravi, N.; and Rudin, C. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models .

Nitzan, Y.; Bermano, A.; Li, Y.; and Cohen-Or, D. 2020. Disentangling in Latent Space by Harnessing a Pretrained Generator .

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874–1883.

Simonyan, K.; and Zisserman, A. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in neural information processing systems* 1.

Singh, A.; and Singh, J. 2016. Super Resolution Applications in Modern Digital Image Processing. *International Journal of Computer Applications* 150(2): 6–8.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Yang, C.; Shen, Y.; and Zhou, B. 2019. Semantic Hierarchy Emerges in Deep Generative Representations for Scene Synthesis .

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric .

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image Super-Resolution Using Very Deep Residual Channel Attention Networks .

Zhu, J.; Shen, Y.; Zhao, D.; and Zhou, B. 2020. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049* .

Zhu, P.; Abdal, R.; Qin, Y.; and Wonka, P. 2019. SEAN: Image Synthesis with Semantic Region-Adaptive Normalization .