ZSRGAN: Zero-shot Super-Resolution with Generative Adversarial Network

Yujin Lee
dldbwls0505@kaist.ac.kr

Samuel Teodoro
sateodoro@kaist.ac.kr

Gihoon Kim
gihoon@kaist.ac.kr

Abstract

Deep learning based super-resolution (SR) is one of the most actively studied areas of computer vision. However, many of these studies are conducted on a supervised manner, requiring a large amount of data. There are several problems with this. First, the ground truth and input pair of the dataset is made using only a specific procedure, usually the bicubic downsampling. As a result, supervised SR works well only for these images, introducing a second problem where the model can not work well for test images not found in the training distribution. Based on this, it is difficult to say that these methods indeed is super-resolution for raw images found in the real world. In this paper, we introduce a novel network named Zero-shot Super-Resolution with Generative Adversarial Network (ZSRGAN) for real world image SR, which needs only one test image and does not rely on any other external datasets. Unlike existing methods, we propose optimization on the perceptual aspect as well as reconstruction of pixel units using zero-shot SR method. Therefore, through our proposed model, a real world image can be super resolved with the best perceptual quality without any information from additional datasets.

1. Introduction

Super-resolution (SR) is one of classical problems in computer vision. SR aims to recover the right high-resolution (HR) image from a low-resolution (LR) image. Single image super-resolution (SISR) is a super-resolution task that super-resolves a single LR image to an HR image. One of the popular SISR methods is the interpolation-based method. This method works such that an LR image is upsampled using a predetermined equation in the local patch. However, since this approach is based on relatively simple linear mapping, it is difficult to capture complex and nonlinear relationships between LR and HR images.

Recently, learning-based methods are attracting attention as technologies that can well express nonlinear relationships between LR and HR images which is the limitation of interpolation-based methods. Since the emergence of convolutional neural networks (CNN), computer vision tasks has improved significantly. Specifically, the conception of SRCNN [5] stimulated numerous SISR research. And one of the most studied aspects is how to improve SR performance based on their Peak Signal-to-Noise Ratio (PSNR) score. Various studies [2, 4, 5, 10, 12, 22] have proposed methods to obtain high PSNR scores, producing *state-of-the-art* models in the process. However, it is evident from these studies that the output images of the PSNR-oriented approach are not realistic. To add, this approach needs a significant amount of LR-HR pairs for training.

To produce not just high scores but also high quality SR images, architectures like the SRGAN [11] and ESRGAN [19] are proposed. These models are based on the Generative Adversarial Network (GAN) paradigm [7]. These GAN-based approaches generate output images that have better perceptual quality than the PSNR-oriented approach. However, this method also requires a lot of training images.

On the other hand, there are also zero-shot-based approaches like the ZSSR [15] and MZSR [17]. Zero-shot learning is a word mainly used in classification works. The term refers to the problem of predicting the categories for inputs which are not in the training data or categories. Hence, ZSSR [15] proposes a method that performs SR with only one LR image by learning its internal features.

Shocher et al.[15] point out that models trained in a supervised manner using huge datasets like DIV2k [1] overfits in a specific downscaling method, the bicubic downsampling. The paper shows that these models perform well only for bicubic-downscaled LR images. Meanwhile, ZSSR [15] produces outputs with better visual quality on non-ideal LR images, i.e., images that are not downscaled through the bicubic method, while previous models perform poorly on these. However, ZSSR [15] has a very simple architecture, so there is a possibility that it can be further improved by conducting more research. And so far, there is no very suitable model that guarantees an SR image with high quality from a real world LR image which have no HR pair and no

information about how it is downscaled.

To solve this problem, we introduce a zero-shot based super-resolution method with GAN paradigm. We name this model as ZSRGAN. For our experiments, we use the dataset provided for MZSR[17] experiment. It has non-ideal downscaled LR images and their HR pairs. Using zero-shot based approach and GAN-based approach simultaneously, ZSRGAN not only performs well for non-ideal LR images but also produces SR outputs which have better quality than previous zero-shot based approaches without any training datasets. Our main contributions are:

1. We propose a novel network named Zero-shot Super-Resolution with Generative Adversarial Network (ZSRGAN) for the first time.

2. Our model generates the most natural images in an unsupervised manner with only one test image without any other external datasets.

3. We provide this highly usable framework as an open source project in the following link: https://github.com/rlgnswk/AI604_Team19.

2. Related Works

2.1. PSNR-Oriented Approach

PSNR-oriented approach is one active research topic in the SR field. PSNR score is commonly used to measure the reconstruction quality of lossy compression. SRCNN [5] is the first CNN-based model to tackle the SR task. It has only 3 convolutional layers but it produced a PSNR score which is quite an improvement from previous methods. Another architecture is the VDSR [10] which is a network that uses skip-connections. It is a very deep network but it did not suffer from the exploding or vanishing gradient problems. In recent research, SR methods started employing the attention mechanism. RCAN [22, 4, 2], for example, applies the residual structure and attention mechanism to adaptively rescale features, achieving *state-of-the-art* PSNR score.

2.2. GAN-based Approach

Generative Adversarial Network (GAN) [7] is a novel architecture for generating image from a latent vector and it consists of a generator and a discriminator. Conditional GAN [13] is designed for generating from input and latent vector together. Generator generate an output image to deceive discriminator, and discriminator determines not to be deceived by generator. GAN has the advantage of creating a realistic image with this adversarial training process.

In general, like in PSNR-oriented approaches, SR networks learn from the pixel-wise difference between their SR output and its corresponding ground truth HR image. However, outputs of these approach appear blurry. To solve this problem, it was proposed to use GAN-based on unsupervised learning instead of the existing supervised learning approach. The SRGAN [11], the first proposed GAN-based approach, used the perceptual loss [9] that combines the loss obtained from the difference between the feature maps of the VGG [16] network and the GAN loss from a discriminator. Through this method, the output of SRGAN [11] shows the higher visual quality.

ESRGAN [19] is an improved version of the SRGAN [11]. ESRGAN [19] used a different basic block named Residual in Residual Dense Block instead of previous Residual Block. Batch normalization is removed in the network like in [10]. This model performed very well that it won first place in the PIRM 2018-SR Challenge [3]. Another method that adapted the GAN-based approach is the RankSRGAN [21] which applied rank-content loss related to metric learning mechanism like in [8].

2.3. Zero-shot based Approach

Most LR images are downsampled through a specific method called bicubic interpolation. ZSSR [15] which is the first paper based on zero-shot approach in SR task tackles this. This paper shows that the existing models have poor performance for LR images that have not been downsampled using the bicubic method. However, even if only LR images are used, ZSSR [15] performs better than existing models. The core of this paper says that the information inside the test image itself is more meaningful than the external information because the LR-HR pair has fractal-like relation. MZSR [17] applied meta-learning to ZSSR [15]. Meta-learning means "learning to learn" for how to learn faster. In [17], MAML [6], which is one of popular meta-learning method, was used. MZSR [17] shows better convergence speed.

Wang et al.[18] show that using the Mean Squared Error (MSE) produces over-smooth outputs while using GAN loss produces images accompanied by artifacts. GAN-based approaches reduce the smoothness of its outputs, generating clearer SR images. That is why GAN-based models like [11, 19] generate perceptually more satisfying results. Also, the zero-shot based models [15, 17] show that the super-resolution task for raw LR image needs zero-shot based approach. However, these models use MSE loss for training their models. Therefore, we propose Zero-Shot Super-Resolution with Generative Adversarial Network (ZSRGAN) which applied the combination of these two approaches to generate the appropriate HR image from its LR input.

3. Method

In [15], the authors successfully show that a simple image-specific network will suffice in encoding the LR-HR relations of single images. Hence, our propsed GAN-

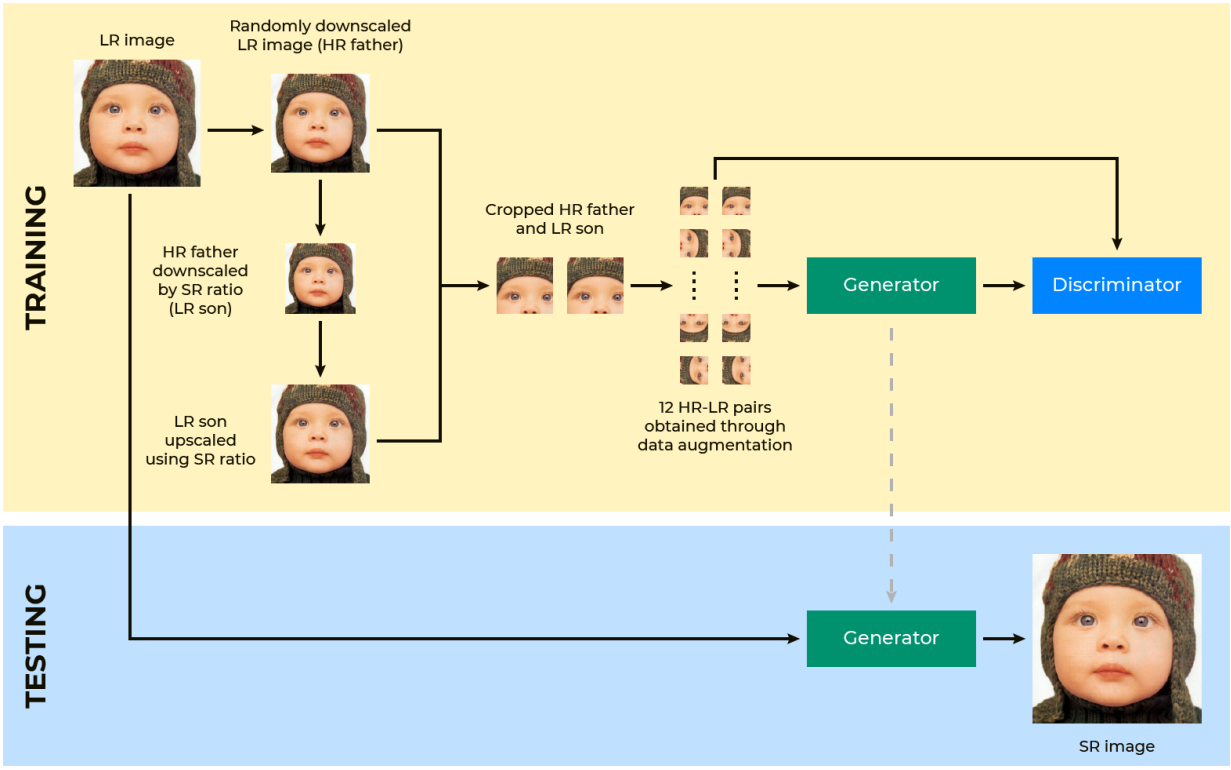


Figure 1: ZSRGAN pipeline. Our proposed method (ZSRGAN) employs a GAN-based approach, where a generator (ZSSR [15]) is trained to synthesize an SR image given LR inputs and the quality of the generated images is assessed by a discriminator. After training, the LR image is upscaled by feeding it to the trained generator.

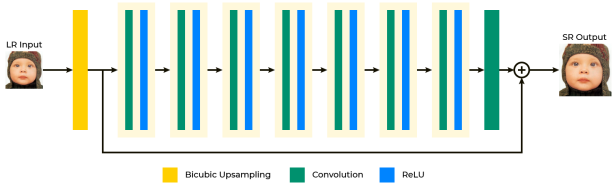


Figure 2: ZSRGAN generator network. We employ the shallow, fully convolutional network of [15] as generator for our approach.

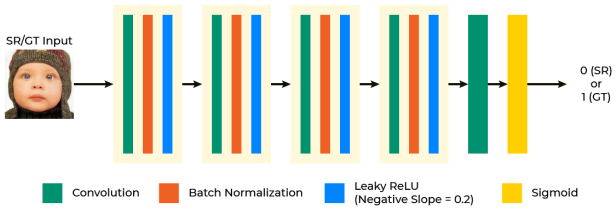
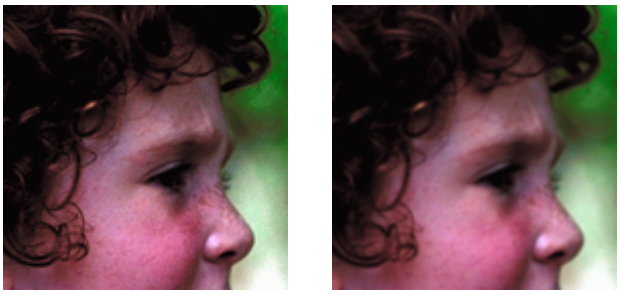


Figure 3: ZSRGAN discriminator network. Our approach utilizes the discriminator of [14] but one convolutional block is removed to reduce the required minimum input image dimension from 128×128 to 64×64.

based approach utilizes small architectures to super-resolve LR images. Figure 1 shows our network pipeline where we train a generator to produce SR outputs from LR inputs and a discriminator judges the quality of the generated SR images. In this section, we give more details about our pipeline, network architectures, and loss functions.

3.1. Pipeline

Our pipeline starts with a single test image I. In order to generate our training data, we rely heavily on how [15] produced theirs. We generate an "HR father" by randomly downscaling I. A list of aspect ratios is first generated where the width of the shorter side could range from 64 pixels up to the original width. We choose an aspect ratio from the list where the ratio closer to the original image dimension has a higher chance of being chosen. After choosing and downscaling the "HR father", we produce the corresponding "LR son" by further downscaling the "father" to the desired SR scaling factor. After this, the LR son is again upscaled using the SR ratio. We use bicubic interpolation for both the upscaling and downscaling stages.



(a) Non-degraded image (b) Degraded image

Figure 4: We used degraded image(right) for non-ideal case. (zoom in for a better view).

The HR father and LR son are then cropped. If the shorter side of the pair is greater than or equal to default size, in our case 128×128, the pair is randomly cropped to this fixed dimension size. Otherwise, we use the width of the shorter side as the dimension of the cropped image. We delay the discussion of our data augmentation method in the next section.

3.2. Generator

The shallow, fully convolutional network of [15, 17] serves as the generator of our GAN-based approach. It consists of 8 convolutional blocks, the first 7 of which contains 64 output channels and followed by a ReLU non-linearity and the last only has 3 output channels with no activation. The kernel size, padding, and stride are 3, 1, and 1, respectively. Finally, we employ a skip-connection [10] between the input and the output images. Figure 2 shows our generator.

3.3. Discriminator

The discriminator of [14] is employed to serve as our network's discriminator. The original architecture contains 5 convolutional blocks (a series of convolution, batch normalization, and leaky ReLU with a negative slope of 0.2), a final convolutional layer, and a sigmoid nonlinearity. With this, the network requires an input image with a minimum dimension of 128×128. However, in our approach, the possible minimum input dimension is 64×64 so we removed one convolutional block, leaving us with only 4. Our first convolutional block outputs 64 channels then doubles in the succeeding blocks until it reaches 512. The final convolutional layer then outputs a single feature map which is fed to the sigmoid activation layer. All our convolutional layers use a kernel size, stride, and padding of 4, 2, and 1, respectively. Our discriminator network is shown in Figure 3.

3.4. Loss Functions

Reconstruction Loss. We use L_1 loss function as reconstruction loss. The reconstruction loss is the content loss that evaluates the 1-norm distance between the output image $G(I^{LR}, z)$ and the ground truth I^{HR}. The reconstruction loss is given by:

$$L_{Rec} = \mathbb{E}||G(I^{LR}, z) - I^{HR}||_1 \tag{1}$$

Perceptual Loss. Perceptual loss compares two different images in the abstracted feature stage using the VGG network. Each GT and output enters the VGG16 network as input. Next, we measure the L_1 distance between the output features of the `relu3_1` layer. This loss can optimize the model not only in the pixel space but also in the feature space of LR-HR pair. The perceptual loss is defined as:

$$L_{Percep} = \mathbb{E}||f_{vgg}(G(I^{LR}, z)) - f_{vgg}(I^{HR})||_1 \tag{2}$$

Adversarial Loss. Our network is image-specific [15], hence, for every test image, we initialize and train a new model. We optimize alternately our generator and discriminator to solve the min-max problem in [7] and [11] given by $G^* = \arg\min_G \max_D L_{GAN}(G, D)$. With this, the discriminator will have a hard time classifying HR images since the generator learns how to create almost real images [11]. The adversarial loss is given by:

$$\begin{aligned} L_{GAN}(G, D) = & \mathop{\mathbb{E}}_{I^{HR} \sim p_{train}(I^{HR})} [log(D_{\theta_D}(I^{HR}))] + \\ & \mathop{\mathbb{E}}_{I^{LR} \sim p_G(I^{LR})} [1 - log(D_{\theta_D}(G_{\theta_G}(I^{LR}, z)))] \end{aligned} \tag{3}$$

Total Loss. In summary, as shown in Equation 4, we use a combination of three losses as generator loss and controlled the ratio among the losses using λ_R, λ_G, and λ_P. The generator loss is given by:

$$\begin{aligned} L_{generator} = \lambda_R * L_{Rec} + \lambda_G * L^G_{GAN} + \\ \lambda_P * L_{Percep}, \end{aligned} \tag{4}$$

where λ_R, λ_G, and λ_P are the tradeoff parameters and are set to $\lambda_R = 1.0$, $\lambda_G = 1.0$, and $\lambda_P = 1.0$.

Additionally, the discriminator loss is given by:

$$L_{discriminator} = 0.5 * L^D_{GAN} \tag{5}$$

We therefore aim to make our model learn the feature space and naturalness of the test image, not just optimization in the pixel space.

(a) Baby from Set5 (b) ZSRGAN (Ours) (c) ZSSR [15]

(d) Input LR (e) Bicubic (f) Ground Truth

Figure 5: Qualitative comparison on the Baby image from the Set5 dataset (zoom in for a better view).

(a) Fish from BSD100 (b) ZSRGAN (Ours) (c) ZSSR [15]

(d) Input LR (e) Bicubic (f) Ground Truth

Figure 6: Qualitative comparison on the Fish image from the BSD100 dataset (zoom in for a better view).

3.5. Evaluation

PSNR. The peak signal-to-noise ratio (PSNR) is an image quality evaluation score. It is mainly used to evaluate image quality by pixel-wise information. This value is based on Mean Square Error (MSE), and the higher the value, the better. As mentioned in Section 2.2 , this method alone is difficult to evaluate perceptual quality, so it was used together with other evaluation method.

$$MSE = \frac{1}{CHW} \sum_{c=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} (x_{c,i,j} - y_{c,i,j})^2 \quad (6)$$

$$PSNR = 10 \log_{10} \frac{MAX_I^2}{MSE} \quad (7)$$

Dataset	Bicubic	ZSSR [15]	ZSRGAN (ours)
Set5	27.1920 / 0.8294 / 0.2164	**27.4566** / **0.8400** / 0.2029	27.1494 / 0.8245 / **0.1991**
BSD100	25.3022 / 0.7319 / 0.3127	**25.5590** / **0.7336** / **0.2828**	25.2476 / 0.7117 / 0.3052

Table 1: Evaluation ZSSR and ours (PSNR↑ / SSIM↑ / LPIPS↓)

SSIM. Structure similarity index measure (SSIM) is the score that evaluate quality of image. It measures the similarity of the original image to the distorted image caused by compression and transformation. Brightness, contrast, and structure are compared between two images.

LPIPS. Learned Perceptual Image Patch Similarity (LPIPS) is proposed by Zhang et al [20]. LPIPS measures the perceptual similarity between the two input patches by calculating the similarity between outputs from various layers of the classification model. In our case, we use LPIPS based on classification model of the VGG network.

4. Experiments

4.1. Dataset

The two datasets provided in [17] are used to evaluate the performance of our model. These datasets contain the images from Set5 and BSD100 with their quality degraded using isotropic Gaussian blur kernel with width $\lambda = 2.0$ followed by direct subsampling kernel. This means that these two datasets are not just bicubic downsampled. Figure 4 shows that an LR image degraded using previously mentioned method is more blurry than the regular LR (bicubic downsampled). The two datasets consist of 5 and 100 LR-HR pairs, respectively.

4.2. Setup

We conducted all experiments by constructing HR-LR pairs using $2\times$ ratio. In the case of our model, the training stage is the test stage. The learning rate is initialized to 1×10^{-3} and the default number of total iterations is 2,000. For every 500 iterations, the learning rate is reduced by half. In addition, Adam optimizer is used for both the generator and discriminator. There is only one batch for each iteration and it consists of 12 HR-LR input pairs.

We perform data augmentation to the HR father and LR son to obtain the input pairs. Specifically, we rotate the the pair using 4 angles of rotation (0°, 90°, 180°, and 270°) and flip horizontally and vertically each rotated pair. Noise, with $\mu = 0$ and $\sigma = 0.1$ as default, is then introduced to the LR images before being fed to our proposed network. After training, we use the original test image I to produce an SR output.

(a) Stones from BSD100 (b) HR

(c) ZSSR[15] (d) ZSRGAN(ours)

Figure 7: Qualitative comparison on the Stones image from the BSD100 dataset: There is an artifact indicated by red arrow. (zoom in for a better view).

4.3. Quantitative Comparison

We compare the performance of ZSRGAN on Set5 and BSD100 datasets. Quantitative comparison using PSNR, SSIM, and LPIPS of the proposed method and existing methods is given in Table 1. Although the PSNR and SSIM is lower than the existing methods, LPIPS is the lowest in our method on the Set5 dataset. This means that, compared to existing methods, the features of the images generated by our network is most similar to the features of the ground truth. In BSD100 dataset, although the LPIPS of our proposed method is not the lowest, it is still lower than the bicubic interpolation. The reason for this poor performance is explained in Sections 4.4 and 4.5

4.4. Qualitative Comparison

Figure 5 and Figure 6 show super resolution images on Set5 and BSD100 datasets, respectively. To qualitatively evaluate the result, we use ZSSR [15] and bicubic interpolation method on each dataset. Overall, the results from each method seem blurry than ground truth image. Also, espe-

cially in the images from ZSSR [15], the boundary between all pixels is clearly visible because it is based on pixel-wise learning and it makes the image as low quality. So, among them, the images from proposed method are natural and realistic due to perceptual loss and adversarial loss. However, because proposed method is GAN-based, there are some artifacts. In Figure 7, some black parts are newly made as an artifact in the image from proposed method.

4.5. Limitations

Even getting image results with better quality, we receive worse LPIPS scores. The cause of this can be found in the problem of the GAN structure. There is some artifact in the detail of output image. This is exactly what Wang et al. [18] argued that the GAN could make some artifact. Therefore, although the output image has more naturally looking, the score may be low because the artifact makes the output image different from GT image. The way to stabilize artifact will be the future works.

5. Conclusion

In this project, we propose a novel SR approach that simultaneously exploits the GAN-based and zero-shot-based single image super resolution frameworks. In this manner, we are able to generate SR images with good perceptual quality without needing a large amount training images. Our experiments show that, although our approach records lower PSNR and SSIM scores than a previous *state-of-the-art* zero-shot-based study, our model can generate more natural-looking results.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 1

[2] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2

[3] Y Blau, R Mechrez, R Timofte, T Michaeli, and L Zelnik-Manor. The pirm challenge on perceptual super resolution, 2018. 2

[4] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 1, 2

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017. 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 2, 4

[8] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015. 2

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2

[10] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 1, 2, 4

[11] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, 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*, pages 4681–4690, 2017. 1, 2, 4

[12] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1

[13] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. 2

[14] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. 3, 4

[15] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018. 1, 2, 3, 4, 5, 6, 7

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[17] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution, 2020. 1, 2, 4, 6

[18] Xintao Wang, Ke Yu, Chao Dong, Xiaoou Tang, and Chen Change Loy. Deep network interpolation for continuous imagery effect transition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1701, 2019. 2, 7

[19] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1, 2

[20] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of

deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[21] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 2

[22] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 1, 2