CoolGAN: GANs with Transformers Super-Resolving Images

Nalin Semwal
Netaji Subhas University of Technology
New Delhi, India
semwalnalin@gmail.com

Abstract

Although Single Image Super-Resolution is an important editing task with many applications, it still remains an ill-posed one. This is due to the existence of a large number of possible HR images for each LR image, and also because popularly used metrics such as PSNR and SSIM promote blurry images that lack realistic detail. In this work, we present CoolGAN, a transformer-based GAN with a focus on generating realistic detail. Our generator consists of Swin transformers acting upon the input image at different levels of detail, and we also use a novel transformer-based perceptual loss to further promote realism. We compare CoolGAN to state-of-the-art methods. Quantitatively, CoolGAN ranks below the others in terms of PSNR and SSIM. However, qualitative analysis reveals that CoolGAN generates far sharper details. We provide some examples here, and more are provided in the supplementary material.

1. Introduction

Single Image Super-Resolution (SISR) aims to obtain a high quality Super-Resolved (SR) image from a degraded Low-Resolution (LR) image. The intention is to deal with blurry images by clarifying, sharpening, and upscaling them. The problem is ill-posed; since the LR images don't possess high-frequency details, there are multiple High-Resolution (HR) images that map to the same LR image upon being downscaled. Some methods try to address this by reformulating the task itself [15], but a majority of SISR methods rely on measurements such as Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) to evaluate their performance. This comes with its own problems - minimizing pixel-wise distance measures between the SR and ground truth HR images leads to better PSNR and SSIM scores, but leads to blurring and lack of detail.

SRCNN [3] pioneered the use of CNNs for SISR; subsequent works such as EDSR [10], DRCN [7], RDN [22], etc. extended this by achieving better PSNR scores. However, these methods lead to over-smoothed results. In order to improve the visual details in the SR image, some methods incorporated perceptual [5] and contextual losses [13]. SRGAN [8] introduced adversarial loss and achieved improved perceptual quality; many methods have since made use of GANs to acheive photorealistic SR images [19, 20].

(a) Bicubic (b) Real-ESRGAN (c) SwinIR (d) CoolGAN (ours)

Figure 1. An Example. The image produced by CoolGAN contains sharper details than the others.

Real-ESRGAN [19], specifically, focuses on practical image restoration and improving the visual quality of the image rather than maximizing PSNR and SSIM. The model is trained with an improved U-Net discriminator to improve realness.

On the other hand, transformers have recently been applied to vision tasks with great success [4, 11]. The transformer architecture is able to exploit global interactions between different regions in the input feature map as opposed

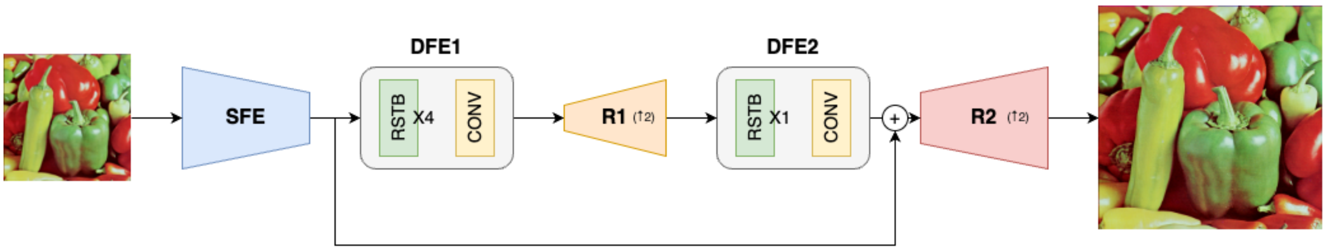


Figure 2. CoolGAN's Generator

to a convolutional layer, whose operation is inherently local. One of the chief drawbacks of vision transformers that limited their application in dense vision tasks was the computational cost, which was addressed in some subsequent work [11]. Following this, vision transformers have successfully been applied in SISR [9, 12].

Swin transformers [11] adapted vision transformers for efficient use in dense vision tasks by means of a shifted window strategy. With this, self-attention is computed locally at each individual module, but the application of shifted windows over multiple layers allows long-range global dependencies to be learned. SwinIR [9] successfully applied these to a SISR generator, achieving state-of-the-art performance in terms of PSNR and SSIM.

We take cue from both of these paradigms to design CoolGAN, a new transformer-based GAN with a focus on realism. For the design of the generator, we incorporate both vision transformers and convolutional processing in the form of an architecture similar to SwinIR [9]; however, our method differs with regards to the level of detail we allow the transformer to act upon. Our generator can be seen as consisting of five stages: 1) Shallow Feature Extraction (SFE), 2) Deep Feature Extraction 1 (DFE1), 3) Resolve 1 (R1), 4) Deep Feature Extraction 2 (DFE2), 5) Resolve 2 (R2). The deep feature extractors DFE1 and DFE2 are where the bulk of the processing takes place; they consist of the Swin transformers in the form of RSTB modules, as presented in SwinIR [9]. The nature of processing done by the Resolve stages differs based on the scale factor. In the case of $\times 2$ and $\times 3$ Super-Resolution, R1 upscales the feature maps to the respective scale and R2 consists only of convolutional refinement. If the scale factor is some other multiple of 2, i.e. $\times 2^n$ Super-Resolution, where $n \in \mathbb{N} - \{1\}$, R1 upscales the feature maps to $\times 2^{\lfloor n/2 \rfloor}$ and R2 performs the final $\times 2^{\lceil n/2 \rceil}$ upscaling to $\times 2^n$. In this preliminary work, we evaluate our approach on $\times 4$ Super-Resolution, wherein both R1 and R2 perform $\times 2$ upscaling. In order to promote perceptual quality, we also design a new perceptual loss based on vision transformers, which combines features from VGG19 [18] and Swin-Tiny [11].

We evaluate CoolGAN on the Set5, Set14 and Urban100 datasets for quantitative comparison with previous methods, and also discuss the perceptual visual qualities of the images produced.

2. CoolGAN

Here we describe the three most important factors of CoolGAN: the network, the generator (as shown in Figure 2), and the perceptual loss function.

2.1. The GAN

The structure of the GAN remains the same as in SRGAN [8]. Given the LR image I^{LR} and the corresponding HR image I^{HR}, our goal is to train the generator G_{θ_G}, where θ_G are the weights parametrizing the network. We seek to optimize θ_G as:

$$\hat{\theta}_G = \underset{\theta_G}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} l^{SR}(G_{\theta_G}(I_n^{LR}), I_n^{HR}) \tag{1}$$

where we define l^{SR} as a weighted sum of three loss components: the pixel loss, the perceptual loss and the adversarial loss. Of these, the pixel loss is the simple L1 loss between the SR and HR image. The adversarial component encourages the generator to try and fool the discriminator, which itself keeps getting better, in a min-max game. We use the relativistic discriminator [6] D_{Ra} as in ESRGAN [20], based on a modified VGG network.

$$l^{SR} = \alpha l_{pixel} + \beta l_{perceptual} + \gamma l_{adversarial} \tag{2}$$

$$l_{pixel} = \frac{1}{rWH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{3}$$

$$\begin{aligned} l_{adversarial} = & -\mathbb{E}_{I^{HR}}[log(1 - D_{Ra}(I^{HR}, G_{\theta_G}(I^{LR})))] \\ & - \mathbb{E}_{G_{\theta_G}(I^{LR})}[log(D_{Ra}(G_{\theta_G}(I^{LR}), I^{HR}))] \end{aligned} \tag{4}$$

The perceptual loss is described in Section 2.3

Method	Set5		Set14		Urban100	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
RCAN [21]	32.63	0.9002	28.87	0.7889	26.82	0.8087
SAN [2]	32.64	0.9003	28.92	0.7888	26.79	0.8068
IGNN [23]	32.57	0.8998	28.85	0.7891	26.84	0.8090
HAN [16]	32.64	0.9002	28.90	0.7890	26.85	0.8094
NLSA [14]	32.59	0.9000	28.87	0.7891	26.96	0.8109
SwinIR [9]	32.72	0.9021	28.94	0.7914	27.07	0.8164
CoolGAN (ours)	30.06	0.8453	26.65	0.7339	24.37	0.7534

Table 1. CoolGAN's PSNR and SSIM as compared to SOTA methods.

2.2. The Transformer

Shallow Feature Extraction (SFE) consists of 3 convolutional layers, with 2 layer normalizations in between. As described in SwinIR [9], convolutional layers are better at early visual processing.

Deep Feature Extraction 1 (DFE1) consists of 4 RSTB blocks (the structure of the RSTB blocks is the same as in SwinIR [9]) with 6 Swin Transformer Layers in each. This module acts at the same dimensions (height and width) as the LR image.

Resolve 1 (R1) upscales the image by $\times 2^{\lfloor n/2 \rfloor}$, where n is the scale factor for the Super-Resolution ($n = 4$ in our experiments) using a PixelShuffle layer [17].

Deep Feature Extraction 2 (DFE2) consists of a single RSTB block. This single block incurs exponentially higher computational cost than the previous RSTB block, which is why we use only one. We use this because this allows global self-attention to be applied at a higher dimension and can recover finer details.

Resolve 2 (R2) upscales the image by another $\times 2^{\lceil n/2 \rceil}$, again using PixelShuffle. This is both preceded and followed by sets of 3 convolutional layers with 2 layer normalizations in between. This module is responsible for the final high-resolution refinement.

2.3. A Better Perceptual Loss

Our perceptual loss consists of 2 components. One is based on VGG features, similar to previous works [9, 20]. For this component, we take the L1 losses between the output feature maps for $G_{\theta_G}(I^{LR})$ and I^{HR} at layers 16, 25 and 34 of the VGG19 network; we sum these using weights 0.5, 0.75 and 0.75 respectively. The second component is based on Swin features, which with their global interactions may be able to better guide the perceptual quality of the image. For this component, we take the L1 losses between the output feature maps for $G_{\theta_G}(I^{LR})$ and I^{HR} at stage 3 of the Swin-Tiny [11] network.

3. Experiments

3.1. Implementation Details

In the RSTB blocks, we use a window size of 8, channel size of 120, and 6 attention heads per MSA module. We train the network for $\times 4$ Super-Resolution with a training patch size of 32 and a batch size of 8 for 250,000 iterations on the DIV2K dataset [1]. We adjust α, β and γ as training goes on. We observed faster convergence in the earlier stages using only the pixel loss (this will be further explored in future work), so we train the network with $\alpha = 1, \beta = 0, \gamma = 0.01$ for the first 100,000 iterations, then with $\alpha = 1, \beta = 1, \gamma = 0.1$. We use the Adam optimizer for both the generator and discriminator, with an initial learning rate of 0.0001, which is halved first at 50,000 iterations and then again at 100,000 iterations.

3.2. Results and Conclusion

For a holistic analysis, we compare our method to various state-of-the-art on PSNR and SSIM scores. These results are presented in Table 1. As we can see, our model scores quite low compared to SOTA. We observed that both the PSNR and SSIM scores decrease after the 100,000 iteration mark, as the perceptual loss is optimized more aggressively.

Following this, we perform a qualitative comparison with SwinIR [9] and Real-ESRGAN [19]. Some illustrative images are shown on the following page. We observe that our model produces images of significantly better perceptual quality; our images are also sharper and contain realistic details. We attribute this to three things: our perceptual loss which contains an extra transformer-based component; our DFE2 module which acts upon a halfway upsampled version of the image; and also our training method as described in Section 3.1. Due to the page limit, more qualitative comparison images are provided in the supplementary material.


(a) Bicubic


(b) Real-ESRGAN


(c) SwinIR


(d) CoolGAN (ours)


(a) Bicubic


(b) Real-ESRGAN

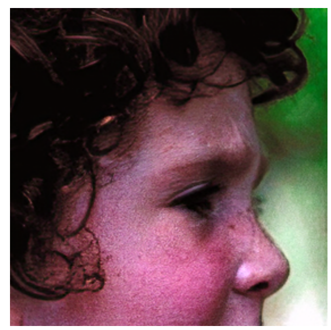

(c) SwinIR


(d) CoolGAN (ours)


(a) Bicubic


(b) Real-ESRGAN


(c) SwinIR


(d) CoolGAN (ours)


(a) Bicubic


(b) Real-ESRGAN


(c) SwinIR


(d) CoolGAN (ours)

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 3

[2] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11057–11066, 2019. 3

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 184–199, Cham, 2014. Springer International Publishing. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 1

[5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 1

[6] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan, 2018. 2

[7] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1637–1645, 2016. 1

[8] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe 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

[9] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 2, 3

[10] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution, 2017. 1

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 1, 2, 3

[12] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution, 2021. 2

[13] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Maintaining natural image statistics with the contextual loss, 2018. 1

[14] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3516–3525, 2021. 3

[15] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models, 2020. 1

[16] 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, 2020. 3

[17] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 3

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 2

[19] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1905–1914, 2021. 1, 3

[20] 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 Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 63–79, Cham, 2019. Springer International Publishing. 1, 2, 3

[21] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks, 2018. 3

[22] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 1

[23] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution, 2020. 3