DCBlock : Efficient Module for Unpaired Image to Image Translation Using GANs

Jin Yong Kim, Myeong Oh Lee, Geun Sik Jo*
Department of Computer Science
Inha University
nastynas9004@gmail.com, eremo2002@naver.com, gsjo@inha.ac.kr

Abstract* — *Recently, as various image-to-image translation studies have been progressed, it is possible to generate high-quality images. In particular, generation models using unpaired data produce meaningful results even building data at a low cost. However, these studies, which are based on Generative Adversarial Networks (GANs), is composed a very heavy architecture. Unlike the commonly used other deep learning models, generally the GANs model consists of two or more in a particular case deep architecture, which has a large computational cost. To solve this limitation, this paper proposes an efficient generator module called DCBlock (Depthwise separable Channel Attention Block). DCBlock consists of a depthwise separable convolution with a relatively low computational cost to replace the standard convolution commonly used in the image to image translation, and channel attention to compensate for information loss caused by depthwise separable convolution. DCBlock showed similar performance to the existing original model while reducing the number of parameters that represents the amount of computation by up to 91.6%. Besides, we experiment with the proposed method for various novel researches and prove that the problem is solved.

Keywords-component *Generative Adversarial Networks , Unpaired Image-to-Image translation, Efficient model architecture, deep learning*

I. Introduction

Recently, image-to-image translation studies using Generative Adversarial Networks (GANs) [1] produce plausible results. GANs can translate the style of the image to another domain [2-5] or generate new high-quality images with high resolution [6,7]. However, GANs are very expensive to compute because of standard convolutional layers, such as convolutional neural networks using very deep architectures (e.g. VGG [8], ResNet [9], AlexNet [10]). Therefore, the number of parameters representing the model's complexity will appear dramatically higher. A large number of parameters have a significant impact on training and inference time and requires high memory resources which is the major limitation for many Image-to-Image translation applications to be applied in real world.

To solve the aforementioned problems, this paper introduces the "Depthwise-separable Channel attention Block (DCBlock)" which replaces standard convolution with depthwise separable convolution and applies channel attention for an efficient unpaired image to image translation. DCBlock dramatically reduces number of trainable parameters that enables use of GANs in applications with limited resources. When we first tried to reduce the number of parameters, we replace standard convolution with depthwise separable convolutions. However, depthwise separable convolution is known to cause information loss [11,12]. Information loss causes poor quality image generation in the GAN model. At this point, we considered how to generate the image as natural as the existing other methods and were inspired by Zhang et al [13] who using channel attention in the residual block. Applying the channel attention focuses on the important parts in feature and make up for information loss, thus ensuring the quality of image. Therefore, we applied the techniques mentioned earlier to create a module called DCBlock. DCBlock is a replacement for "Resblock" [9] which is usually used in GAN architectures [2-5] for image-to-image translation.

Overall, our contributions are as follows:

- We propose a DCBlock that reduces number of parameters and generate almost similar quality images as existing image-to-image translation models
- We have demonstrated how to use channel attention to avoid information loss in depthwise separable convolution.
- We provide experimental results, including quantitative and qualitative assessments of our results with existing models and ablation study on the effect of channel attention on our model

DOI reference number: 10.18293/SEKE2020-028.
*Corresponding Author

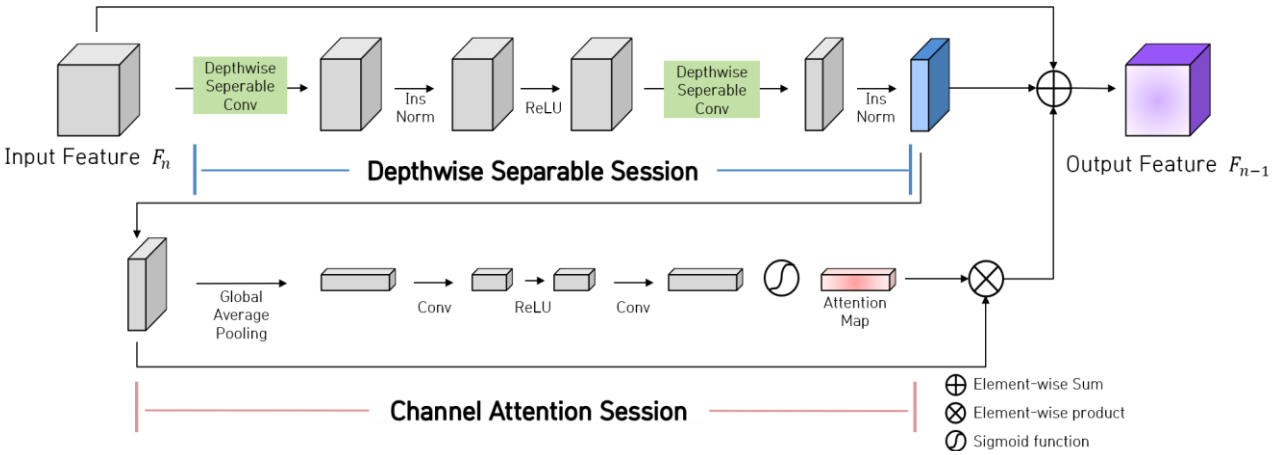

Fig. 1 . DCBlock Architecture

In conclusion, our method can generate the same quality image even though we reduce the parameters of existing baseline models.

II. Related Work

Generative Adversarial Networks. Generative Adversarial Networks (GANs) have shown great performance in image generation and image translation [2-5]. Inspecting image generation mechanism of GANs, generator tries to produce fake images that are indistinguishable from the real ones, while a discriminator tries to distinguish the real image from the fake or generated images. Since two networks are opposing, what each network learns is called "adversarial loss" which is a key point in GAN. In the basic GAN [1] model, there is one generator and one discriminator, but nowadays there are many models with multiple generators and discriminators depending on purpose.

Unpaired Image-to-Image Translation. Image translation that proceeds with paired data is often difficult to apply because data is rarely paired in the real world. On the contrary, using unpaired training data is suitable for real-world application, Consequently, there are various GAN methods presented. CycleGAN [2] learns the cycle consistency loss by mapping the two domains separately in two generators. Among the methods using attention guided, AttentionGAN [4] is a model that adds attention mechanism to CycleGAN. It can make the important part changes via the built-in attention mechanism without the need for additional labeled data or models. The case of multimodal is more efficient than the above models when generating a diversity of images. StarGAN [3] consists of one generator and several discriminators, which efficiently generate various images to increase efficiency and quality. MUNIT [5] creates multimodal images without any guides. MUNIT is composed of contents encoder and style encoder for recombine random noise with input contents at style space.

Efficient CNN Architecture. Many networks using depthwise separable convolution have been studied for efficient neural networks. At first, Xception [11], an architecture inspired by inception and depthwise separable convolution, proposed an extreme version inception module that does 1x1 conv first and then performs spatial correlation mapping on all output channels individually. MobileNet [12] also use depthwise separable convolution and additionally proposed shrinking hyper-parameter consisting of a width multiplier to control the input and output channels and a resolution multiplier to adjust the size of the input image. ShuffleNet [14] highlights that pointwise convolution is still a high cost area. To solve this problem, ShuffleNet designed channel sparse without connecting all weights, and shuffled groups to prevent the problem of getting only information flow for a specific area as input.

As we have seen, efficient CNN networks are being actively researched and real-world applications using them are actively being developed. Therefore, we will introduce a module to be used in the GAN method for efficient image-to-image translation.

III. DCBlock

To address the heavyweight model that unpaired image-to-image translation with GAN has, we proposed DCBlock (Depthwise separable Channel attention Block). At first, we applied depthwise separable convolution to reduce number of parameters. However, as can be seen in Xception [11], it causes information loss. Xception bridges this gap with residual connection but in image-to-image translation did not alleviate it, causing poor generation. Since depthwise separable convolution is performed for each channel, the loss of the feature appearing in the whole part is inevitable.

Accordingly, to tackle an optimal balance between the quality of output and computational cost, we had to add a technique to compensate for information loss. Therefore, we applied an attention module to keep the information we need as much as possible and not lose it even in deep architectures. Our proposed module DCBlock is shown in **Fig 1**. It consists

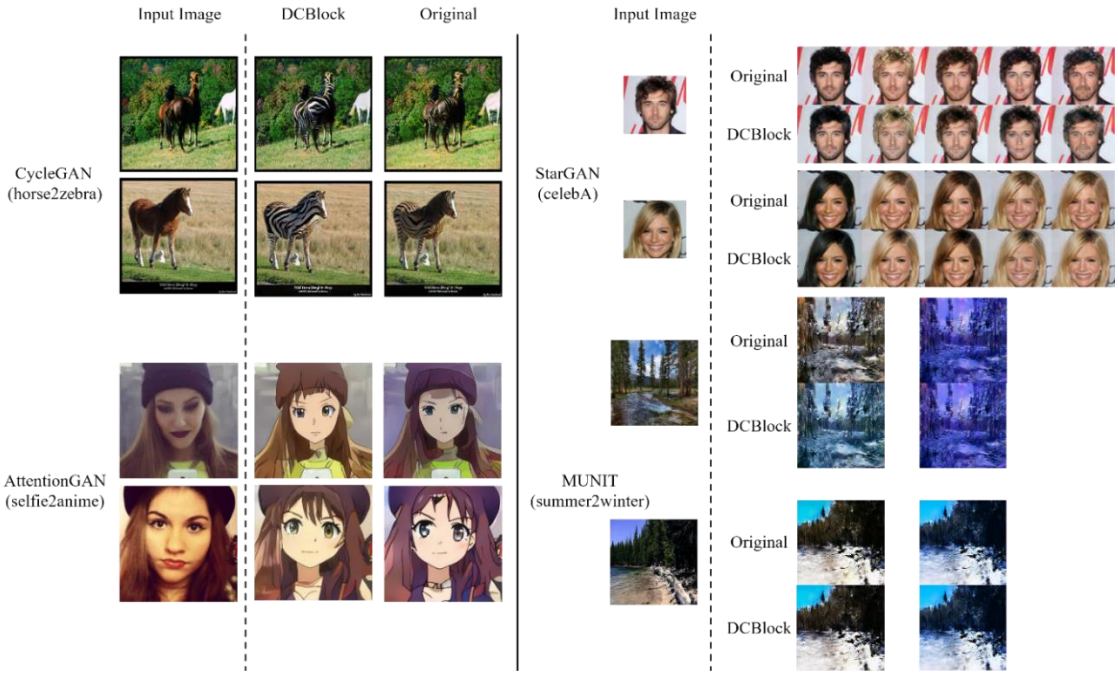

Fig. 2. Output Comparison of The Two Methods

of two sessions, Depthwise separable session and Channel attention session. In DCBlock, the input feature and output feature will be feed to each block proceeding residual learning. For input feature F_n, DCBlock can be formulated as follows:

$$DC(F_n) = F_n \oplus S_{depth}(F_n) \oplus CA(S_{depth}(F_n)), \quad (3)$$

where $DC(\cdot)$ denote DCBlock module and $S_{depth}(\cdot)$, $CA(\cdot)$ denote depthwise separable session and channel attention session. Depthwise separable session significantly reduces the number of parameters than the standard convolution of the existing resblock. And channel attention that denoted $CA(S_{depth}(F_n))$ is formulated as,

$$CA\left(S_{depth}(F_n)\right) = S_{depth}(F_n) \otimes S_{att}\left(S_{depth}(F_n)\right), \quad (4)$$

where $S_{att}(F_n)$ is attention map extracted by channel attention. This is a statistic of the channel obtained through the gating mechanism [15], which can prevent the information loss. In summary, we present DCBlock which a method that is efficient and produces quality similar to existing models. In the next part, we describe a detailed description of the DCBlock configuration.

Depthwise Separable Convolution. As explained in the previous part, to reduce number of parameters, we used depthwise separable convolution, which composed a combination of depthwise convolution and pointwise convolution. In depthwise convolution, there are filters for the number of channels to extract spatial features, which is why the number of input and output channels is the same. Depthwise convolution is can be written as,

$$\hat{G}_{k,l,m} = \sum_{i,j} \hat{K}_{i,j,m} \times F_{k+i-1,l+j-1,m}, \quad (5)$$

where $\hat{K}$ denote kernel size of depthwise convolution, i,j,m denote width, height, input channel and F denote feature map. And pointwise convolution is a 1x1 convolution, and the size of the filter is fixed to 1x1. In contrast to the depthwise convolution, pointwise convolution is performed only on the channel without dealing with spatial features. This helps to greatly reduce the amount of computation in DCBlock. **Table I** shows the differences between the parameters and computational costs of the two convolutions.

Table I
COMPARISON OF TWO KIND OF CONVOLUTIONS

Method	Standard Convolution	Depthwise Separable Convolution
# of Param	$K^2 \times C \times M$	$C \times (K^2 + M)$
Computational cost	$K^2 \times C \times M \times H \times W$	$C \times H \times W \times (K^2 + M)$

where K denote kernel size, C, M denote input and output channel size and H, W denote input height, width. As **Table I** shown, Standard convolution has a computational cost of $K^2 \times C \times M \times H \times W$ while depthwise separable convolution has $C \times H \times W \times (K^2 + M)$. Dividing the two costs to see the difference shows that the cost has reduced by $\frac{1}{M} + \frac{1}{K^2}$. This has great effect in reducing proportionally increasing parameters in GANs where relatively deep networks are used.

TABLE II
QUANTITATIVE EVALUATION

Method		Generator Million Parameter	LPIPS↑	FID ↓	SSIM↑	IS↑
CycleGAN	ResBlock	11.06	Evaluate only multimodal	210.48	0.7898	1.3771
	DCBlock (Ours)	0.89		195.04	0.8495	1.4354
AttnetionGAN	ResBlock	11.82		221.09	0.4392	1.5045
	DCBlock	3.62		226.16	0.4166	1.4620
StarGAN	ResBlock	8.43	0.114	17.61	0.8221	3.2183
	DCBlock	2.56	0.109	19.36	0.8252	3.1949
MUNIT	ResBlock	15.02	0.047	105.94	0.3344	1.8457
	DCBlock	7.22	0.044	105.81	0.3348	1.8322

Channel Attention Mechanism. Channel attention, proposed by Zhang et al [13], is one of the attention's variations that leverages the interdependencies between the channel to focus on informative feature. It uses global average pooling to compress channel information and restore the feature through the convolution layer. Subsequently, the channel statistics are extracted via the gating mechanism [15]. This effect enhances and restores the feature for the focused part of the network. Consequently, we use channel attention to sustain as many features as possible even in deep architectures and to address information loss caused by depthwise separable convolution. We provided more details about Channel Attention Mechanism usage in the next section and show how it affects to generated images in section 4.

IV. Experiments

To explore the suitability of proposed model, we evaluated DCBlock quantitatively and qualitatively on various datasets, comparing different models. The method of experiments is replacement of the "Resblock" [9] with a "DCBlock" on other novel models as we mentioned at section II.

Baseline Models. As baseline models, we adopt CycleGAN[2], AttentionGAN [4], StarGAN [3] and MUNIT [5]. Since they include resblock in their models, they are suitable models for evaluation. For comparison, we apply DCBlock to aforementioned models, and compared the performance and number of parameters. **CycleGAN** consists of two generators and two discriminators, where the generator typically uses nine resblocks, which could be six or U-Net [16] depending on the resolution of images dataset. We used horse2zebra datasets [2] for this task. **AttentionGAN** has a similar architecture to CycleGAN. It has also nine or six resblocks and additionally produce attention mask via a built-in attention mechanism. For the experiment, we used the selfie2anime dataset created in Kim et al [17] for the AttentionGAN. **StarGAN** consists of one generator and N discriminators, where N is datasets number of class. In addition, StarGAN generator consists of six resblocks. For StarGAN experiments we used celebA dataset. **MUNIT** consists of a content encoder and a style encoder as described in the paper. In the MUNIT model, resblocks were applied only for the content encoder. For the experiments, we use summer2winter yosemete dataset [2] with MUNIT model.

Evaluation Metrics. In quantitative evaluation, we measured the quality and diversity of the image with four metrics along the baseline papers. **FID** [18] uses the inception-v3[22] model to extract features and measure the distance between the distribution of real and fake images. The lower value of FID is more similar with real image. **LPIPS** [19] is perceptual metric about patch images. LPIPS evaluate the distance between images patches. Although It indicated that higher is different, and lower is more similar, we use it as a measure of the diversity of the image. **SSIM** [20] is a metric that handles the structure of an image. It evaluate three factors that affect perceptual quality: average brightness, contrast, and structure. SSIM indicates higher is similar. **Inception Score (IS)** [21] is similar to FID in that it uses Inception network [22]. However, they differ in the way they use features, and IS evaluates how diverse the image is and how well it can be determined. In qualitative evaluation was conducted perceptual study according to Kim et al [17]. We conducted a user study in which users voted on their preference image.

Quantitative Evaluation. As seen in **TABLE II**, we applied DCBlock to adopted models. In CycleGAN, replacing the generator's resblock with a DCBlock shows that the FID, SSIM, and IS are similar or better, despite a 91.6% reduction in the number of parameters from 11.06M to 0.87M. In AttentionGAN, SSIM and IS were lower than original model when applied, but FID was higher, and the number of parameters decreased by 69.3% from 11.25M to 3.3M. In

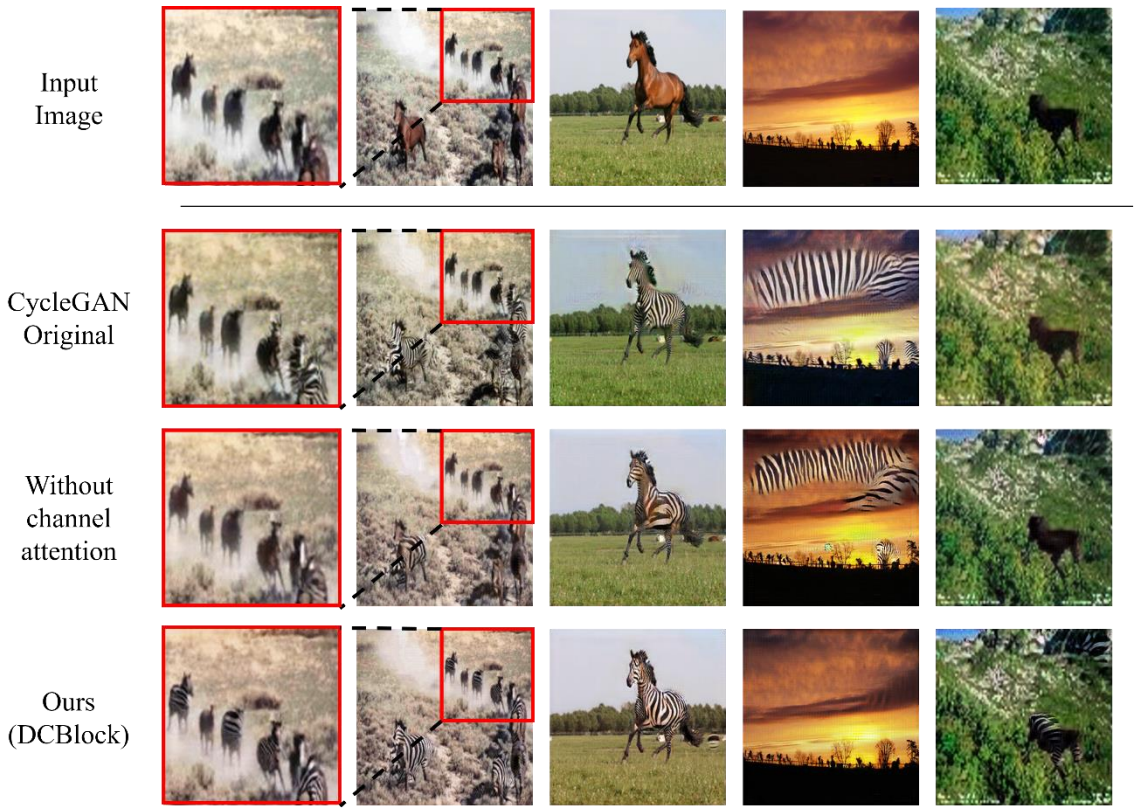

Fig. 3. Ablation Study Comparison

StarGAN, number of parameters reduced by 70.4% from 8.43 to 2.5, and other metric show that original and DCBlock generate images of similar quality. And in result of MUNIT, the number of parameters decreased by 51.9% from 15.02 to 7.22, and similar figures in other metrics as well.

Analyzing the results of quantitative evaluation, we can see that DCBlock shows similar performance as resblock, a method of each model, while dramatically reducing the number of parameters.

Table III
PREFERENCE PERCENTAGE OF USER SCORE

Model	Original (Resblock)	DCBlock(Ours)
CycleGAN (horse2zebra)	51.4(257)	48.6(243)
StarGAN (CelebA)	54.2(271)	45.8(229)

Qualitative Evaluation. We provide 10 randomly sampled images from CycleGAN [2] (horse2zebra) and StarGAN [4] (CelebA) with DCBlock and the original model pairs to 50 participants, and participants evaluate the fake images to choose what they think is more natural. We inform the participant that only the domain of fake image and the original image. The results of the user study are shown in **Table III**. As we aimed for, both methods showed nearly similar preference percentages. First, the two methods of CycleGAN have a difference of 2.8% p (14 votes), which is higher than our method. And StarGAN showed the original with an 8.4% p (42 votes) high preference.

As a result, as shown in **Fig. 2**, our module produces a very similar level of image quality, although it is slightly less in terms of preference than the original module.

Ablation Study. As discussed in section 3, channel attention [13] is an important contribution to image quality in this paper. Our ablation study examines how channel attention affects our module and contribute to generating images. To verify the impact of channel attention, we compared CycleGAN's original model, the "without-channel-attention" model using only depthwise separable convolution, and DCBlock. In terms of number of parameters, CycleGAN, as provided **TABLE IV**, has 11.06 million parameters, without-channel-attention has 0.66 million parameters and ours has 0.89 million. In **Fig 3**, several failure cases are shown in CycleGAN original model and without channel attention module. Second (The enlarged images are the first column) and fifth column images were not translated to zebra when horses in blurry form were used with CycleGAN and without-channel-attention model. However, when we use our module, we can see that the blurry horses are also translated to zebra. And as shown in fourth column, input image is difficult to recognize the horse on the hill with human perception. Despite CycleGAN and Without-channel-attention failure that translate wrong part (i.e. sky), our

method overcomes this issue. As a result, adding channel attention increases a small number of parameters, however, contributed to generate the similar quality as the existing model or improve to better results. This is clear to say that channel attention contributes to extracting the righter focus even in the images that are difficult to distinguish.

Table IV
COMPARISON OF ABLATION STUDY

Model	Millions of parameters
CycleGAN original	11.06 M
Without-CA	0.66 M
Ours (DCBlock)	0.89 M

V. Conclusion

In this research, we proposed DCBlock that solves high computational cost problem in the unpaired image-to-image translation with GANs. DCBlock overcome the large number of parameters and high requirements of memory resources while ensuring the quality of the image. Experimental results show that when our method is applied to the baseline method, it generates images of similar quality or more natural to the existing method while reducing the number of parameters. Since DCBlock is a lightweight network, it is thought to be easier to use in a real world with limited resources.

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-0-01642) supervised by the IITP(Institute for Information & communications Technology Promotion)

Reference

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems (NIPS), 2014, pages 2672–2680.

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[3] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797 .

[4] H. Tang, H. Liu, D. Xu, P. Torr, N. Sebe. AttentionGAN: Unpaired Image-to-Image Translation using Attention-Guided generative Adversarial Networks. arxiv preprint. Arxiv : 1911.11897, 2019

[5] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in European Conference on Computer Vision (ECCV), 2018, pp. 172–189.

[6] C. Ledig, L. Theis, F. Husz´ar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photorealistic single image super-resolution using a generative adversarial network," in IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.

[7] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in European Conference on Computer Vision, 2018, pp. 63–79.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 2012, pages 1097–1105.

[11] Chollet, F. Xception: Deep learning with depthwise separable convolutions. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 1800–1807

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko,W.Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[13] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in European Computer Vision Conference (ECCV), 2018.

[14] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.

[15] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv :1709.01507 (2017)

[16] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234{241. Springer (2015)

[17] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," arXiv preprint arXiv: 1907.10830, 2019.

[18] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in NIPS, 2017.

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.

[20] 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, vol. 13, 2004.

[21] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in Neural Information Processing Systems, pp. 2234–2242, 2016.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826, 2016.