Evaluating the Impact of Self-Attention in Pix2Pix for Image-to-Image Translation

Zheng Liao

School of Computer Science and Network Engineering, Guangzhou University, Guangzhou, China
32306300021@e.gzhu.edu.cn

Abstract. In this work, a self-attention module is incorporated into the generator of the Pix2Pix model and its effects on image-to-image translation with Facades dataset is assessed. The proposed architecture adds self-attention at the bottleneck of the U-Net generator to capture global context while retaining the original generator–discriminator structure. A detailed assessment, including training loss curves, discriminator dynamics, qualitative image comparison and Fréchet Inception Distance (FID), was performed to study the influence of attention on the perceptual output quality and on the optimization behavior. The experimental results demonstrate that including self-attention leads to more stable adversarial loss curves, a lower and more stable L1 reconstruction loss, and more balanced discriminator responses, indicating dynamics of training that are different from that of the vanilla Deep Convolutional Generative Adversarial Network (DCGAN).However, this modification does not produce the perceptual faithfulness benefits: the synthesized images are still visually on par with those produced by the baseline Pix2Pix model and the FID score does not display a visible drop. These results show that at least for the considered Facades dataset and the current experimental training setup, the benefits of the self-attention module manifest more in the way of training stability than in perceptual quality improvements of the generated images. The work provides empirical understanding of attentional mechanisms in conditional GANs as well as suggestions for further research such as multi-level attention, perceptual loss integration, and evaluation on more challenging datasets.

Keywords: GAN, Pix2Pix, Attention Mechanism, Self-Attention.

1 Introduction

Generative Adversarial Networks (GANs) and its variants have shown promising results and the generated images are often with high visual fidelity [1]. For instance, in the area of conditional generation, scientists add class or semantic information to the generation process, which allows the model to create images that meet certain semantic or category specifications. Conditional GANs (cGANs) are a sample of such work, which brings in conditional labels in both the generator and discriminator networks, to obtain control over the generation process [2].

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_83

Beyond supervised conditional generation, unsupervised image-to-image translation has also gained significant attention, with CycleGAN being a landmark work [3]. CycleGAN achieves unpaired image translation between two domains via cycle consistency loss, eliminating the need for aligned training data and expanding the applicability of GAN-based translation models. While it does not rely on explicit conditional labels as traditional cGANs do, it implicitly leverages domain information to guide the generation process, complementing the supervised paradigm of cGANs.

Subsequently, the Pix2Pix model was proposed as a classical image-to-image translation framework based on conditional generative adversarial networks (cGANs) [4]. Pix2Pix incorporates conditional image information into the generator input, allowing the model to learn a mapping from the input image domain to the target image domain, thus enabling tasks such as edges-to-photo, grayscale-to-color, and semantic segmentation-to-real image translation. Its generator adopts a U-Net encoder–decoder structure to capture both local and global image features [5], while the discriminator follows a PatchGAN design to evaluate local realism—a design choice consistent with early convolutional GAN architectures such as DCGAN [6].

However, GANs still face numerous challenges, such as unstable training, severe mode collapse, and difficulty in balancing diversity and quality of generated images [1]. To this end, researchers have investigated a number of model architecture, training strategy and loss function improvements [7, 8]. With respect to the improvements in model architecture for conditional generation, Pix2PixHD is a significant upgrade of the original Pix2Pix [9]. This model is specially tailored for producing high-resolution conditional GAN images using multi-scale generators and discriminators, which not only synthesizes 2048×1024 high-resolution images with significantly improved visual quality but also substantially elevates semantic consistency and the details of synthesized images——solving the issue of the original Pix2Pix failing to generate high resolution images. For long-range dependency modeling, key technical requirement for this high-resolution synthesis task, theoretical foundation was given by Transformer [10]. Concurrently to this work, parallel research was demonstrated on how to include self-attention, without any local-inductive biases, as a fundamental operator to model global contextual relations in GANs, which proved to be a core architectural element for adapting attention-based advancements within GANs.This key insight from Transformer inspired novel work on attention mechanisms that have been developed as a powerful technique to tackle the challenge of GANs – for instance for capturing long-range dependencies crucial for high-resolution synthesis – as they let networks focus attention on salient input features that encompass spatially wide-ranged areas. Their adaptation to GANs has been especially impactful in the quality of feature representation since it allows generators and discriminators to break free from local convolutional restrictions. By applying differentiated weights over spatially or semantically important regions, these modules capture long-range dependencies, increase structural consistency, and enhance the modeling of global context – establishing a key building block to push the frontier in conditional generation tasks such as high-resolution image synthesis. For instance, Self-Attention Generative Adversarial Networks (SAGAN) were one of the first pieces of work to introduce self-attention modules to GANs, explicitly based on the core mechanism of Transformer

[11]. These blocks allow bipolar generators or discriminators to produce "attention" weights across different spatial locations of the feature maps, and therefore these modules explicitly account for long-range relation between pixels or features. In line with the bottleneck principle used in the Transformer-based architecture, this strategy is normally performed on the middle bottleneck layer for generator or discriminator which is most affected by the, integration of global features helps to enhance the integrity of the generated features.

To investigate the effects and potential applications of attention mechanisms in conditional generative adversarial networks, this work introduces a self-attention module into the generator of the classical Pix2Pix model [3]. In this model, the generator still adopts a U-Net encoder–decoder architecture to maintain efficient modeling of both local and global features [5]. Unlike the conventional Pix2Pix, a self-attention module is embedded in the U-Net bottleneck layer, allowing the generator to establish long-range dependencies in the feature space, thereby investigating the potential impact of attention on global feature modeling and contextual correlations. Furthermore, this design retains the skip connections of the original U-Net, providing a structural basis for the interaction between low-level features and high-level semantic information. This study focuses on analyzing the feature representation capacity, training behavior, and the influence on the generated images under this modified structure, providing insights for more effective integration of attention mechanisms in conditional generation models.

2 Method

2.1 Overall Approach

The paper is based on the Pix2Pix architecture (generator-discriminator adversarial learning) and adopts self-attention mechanism in the bottleneck of U-Net generator. And the purpose is to examine whether it can learn long-range dependencies over spatial positions. It should be noted that, as the computational complexity of self-attention is O((HW)2), the consine attention is computed only at the lowest resolution bottleneck feature maps in order to lessen the extra computational and memory cost while preserving modeling global context.

The paper conducts experimental evaluations on the Facades dataset to demonstrate the efficacy of our proposed architecture. The Facades dataset is a popular benchmark used in image-to-image translation for testing the model's ability to translate building facade labels to realistic images of buildings [12]. This dataset is an ideal case to test whether the attention mechanism can improve global structural consistency and translation quality.

2.2 Baseline: Pix2Pix Network

Generator: The generator follows the original U-Net structure. The encoder progressively downsamples the input (e.g., for 256×256 images, num_downs = 8), and the decoder progressively upsamples and concatenates features from the corresponding

encoder layers via skip connections. The final output is generated through a 7×7 convolution followed by Tanh activation.

Discriminator: A PatchGAN discriminator is used, with a 4×4 kernel, stride 2, and padding 1. The number of channels doubles at each layer, and the output is a single-channel patch-level real/fake score map.

Normalization & Initialization: Normalization layers support BatchNorm, SyncBatchNorm, InstanceNorm, or None. Weight initialization methods include normal, Xavier, Kaiming, and orthogonal, with default normal initialization (variance 0.02), consistent with Pix2Pix.

2.3 Self-Attention Module

The paper adopts a 2D self-attention mechanism similar to SAGAN. Given an input feature map of size B×C×H×W:

1.Generate query Q and key K via 1×1 convolutions with channel reduction C→ C/8, and value V with channel C.

2.Compute the attention matrix $A = \mathrm{softmax}(Q^{T}K) \in R^{B\times N\times N}$, where $N = H \times W$.

3.Aggregate features as $O = V \cdot A^{T}$, reshape to $B \times C \times H \times W$, and produce the final output $y = \gamma \cdot O + x$, with γ as a learnable scalar initialized to 0.

This design allows each spatial position to reweight features based on correlations across the entire feature map. Optional channel, spatial, and CBAM attention blocks are also implemented for potential ablation studies, but the main experiments use only self-attention.

2.4 U-Net Generator with Attention

The self-attention module is inserted at the bottleneck of the U-Net generator:

The innermost U-Net block produces output with doubled channels due to skip concatenation.

Self-attention operates on this bottleneck feature (highest-level, lowest-resolution) to capture long-range dependencies.

The rest of the decoder follows the standard U-Net structure, gradually upsampling to the output resolution.

2.5 Training Details

Experiments are conducted on the Facades dataset under the aligned image-to-image translation setting. Each image pair contains a building façade label map and its corresponding real photograph. All images are resized to 286×286 and then randomly cropped to 256×256, with horizontal flipping applied during training. Each epoch consists of 400 training iterations, following the standard Pix2Pix training protocol.

Two models are evaluated in this study:

Baseline Pix2Pix:

The baseline model uses the standard U-Net generator with 8 downsampling and 8 upsampling blocks, and a 70×70 PatchGAN discriminator.

Attention-Pix2Pix (Proposed Model):

The proposed model keeps the same generator–discriminator architecture but inserts a self-attention module at the bottleneck layer of the U-Net generator to enhance global context modeling.

Both models are trained using a batch size of 1 and the Adam optimizer with β_1=0.5 and β_2=0.999. The initial learning rate is set to 0.0002, kept constant for the first 100 epochs, and then linearly decayed to zero for the following 100 epochs. Weight initialization uses the normal scheme with a gain of 0.02, and Batch Normalization is applied throughout the networks. A fixed random seed (2025) is used to ensure reproducibility.

The training objective consists of the vanilla GAN loss and an L1 reconstruction loss with a weight of λ_L1=100. During training, we continuously monitor the generator's adversarial loss (G_GAN), the pixel-level reconstruction loss (G_L1), the discriminator outputs for real and fake samples (D_real and D_fake), as well as the overall generator loss (G_total). To quantitatively evaluate the image quality, the Fr é chet Inception Distance (FID) is calculated at the end of every epoch with validation set, allowing a quantitative comparison between the baseline Pix2Pix model and the attention-augmented version proposed by the authors [13].

All experiments are carried out on an NVIDIA GeForce RTX 4080 (12 GB VRAM) equipped workstation. The training is well-fitting in the memory without the need of applying gradient checkpointing, or mixed-precision methods.

3 Result

3.1 Loss Curve Analysis

The loss curves during the training for the baseline Pix2Pix and Attention-Pix2Pix models were thoroughly logged and are presented, with respect to the following five loss terms: generator adversarial loss (G_GAN), generator L1 loss (G_L1), score of discriminator on real samples (D_real), score of discriminator on fake samples (D_fake).The results indicate possible differences in optimization behavior due to the insertion of the self-attention module.

L1 Reconstruction Loss (G_L1) The Attention-Pix2Pix model has the tendency to exhibit smaller values of G_L1, which can be a sign of a particular pixel-level reconstruction behavior, higher but unstable G_L1 values during the whole training, while the baseline Pix2Pix is characterized by significant G_L1 fluctuations accompanied by even higher peaks:

Baseline Pix2Pix: Substantial peaks occur in the mid-training phase. Extracted from the pre-improvement log, representative values include G_L1 = 43.306 at Epoch 5, Iteration 300; G_L1 = 52.700 at Epoch 10, Iteration 400; and G_L1 = 55.288 at Epoch 15, Iteration 300. The overall value range is as wide as 20–55, with high fluctuation frequency. Statistically, G_L1 exceeds 40 in 38 iterations over 200 epochs.

Attention-Pix2Pix: Minor high values appear in a few iterations, such as G_L1 = 44.973 at Epoch 5, Iteration 300 and G_L1 = 47.947 at Epoch 10, Iteration 400. However, the entire value interval is squeezed into 20–45 the peak durations are short and low frequency. The results indicate that G_L1 surpasses 40 only 19 times in 200 epochs, and the mean G_L1 value is around 15% lower than the baseline.

This discrepancy may be due to the fact that the model had access to wider context, but more investigation is needed to verify this claim.

Adversarial Loss (G_GAN) The self-attention model demonstrates a trend of having smoother adversarial loss curves, which may alleviate some oscillatory behavior in the baseline model:

Baseline Pix2Pix: Severe fluctuations occur in the mid-training phase. Extracted from the pre-improvement log, typical values include G_GAN = 4.352 at Epoch 7, Iteration 400; G_GAN = 2.930 at Epoch 18, Iteration 300; and G_GAN = 4.346 at Epoch 32, Iteration 200. During Epochs 20 - 50, G_GAN's maximum fluctuation amplitude (max value - min value) reaches 3.5, which indicates the change in interaction pace of the generator and discriminator is so chaotic.

Attention-Pix2Pix: The G_GAN curve shows significantly improved smoothness, with a nearly 40% reduction in fluctuation amplitude. For example, during Epochs 50–100, the G_GAN values of the improved model remain within the range of 1.5–3.0, with a maximum fluctuation amplitude of only 1.5. Even in the early training phase (Epochs 5–10), there are no "sharp rises and falls" as observed in the baseline, resulting in a more gradual overall optimization process.

These observations might be associated with changes in global feature representation, although this cannot be conclusively determined from loss curves alone

Discriminator Dynamics (D_real & D_fake) The attention-augmented model tends to display more balanced behavior between D_real and D_fake when compared with the baseline model (Table 1):

Baseline Pix2Pix: Extreme deviations between D_real and D_fake are common. Extracted from the pre-improvement log, examples include D_real = 0.013 and D_fake = 0.908 at Epoch 10, Iteration 400; and D_real = 2.546 and D_fake = 0.112 at Epoch 17, Iteration 200. Statistics show that the average difference between D_real and D_fake over 200 epochs reaches 0.8 ± 0.3, suggesting that the discriminator and generator may exhibit periods of imbalance during training.

Attention-Pix2Pix: A balanced state of D_real ≈ D_fake is maintained in most training phases. For example, during Epochs 50 – 100, the average D_real is 0.32 and the average D_fake is 0.29, with a difference of only 0.03. Even in the early training phase (Epochs 10–20), the maximum difference is only 0.6 (D_real = 0.010 and D_fake = 1.054 at Epoch 15, Iteration 100), and balance is quickly restored. Over 200 epochs, the average difference between D_real and D_fake of the improved model is only 0.3 ± 0.1, a 62.5% reduction compared to the baseline.

The observed average G_L1 value in the attention model is approximately 15% lower, with fewer high-value iterations, the G_GAN curve appears smoother, and the average D_real–D_fake difference is smaller.

Overall, these findings suggest that the attention model may exhibit differences in reconstruction behavior, loss stability, and discriminator–generator dynamics compared with the baseline.

Table 1. Quantitative Comparison of Key Loss Metrics Between Baseline Pix2Pix and Attention-Pix2Pix

Metric	Baseline Pix2Pix	Attention-Pix2Pix	Improvement Ratio
Average G_L1	32.6 ± 5.8	27.7 ± 3.2	15.0%
Number of Iterations with G_L1 > 40	38	19	50.0%
G_GAN Standard Deviation	1.8	1.1	38.9%
Average D_diff (D_real - D_fake)	0.8 ± 0.3	0.3 ± 0.1	62.5%

3.2 Image Analysis

The paper compares the final-epoch image triplets (input label real_A, generated image fake_B, ground truth real_B) produced by the baseline and attention-enhanced models.In general, both models render images that have comparable global structure and color consistency. By the comparison of Figures 1-3 and Figures 4-6, the attention-augmented model exhibits noticeable small differences in some local regions, e.g., a bit sharper borders, a bit cleaner thin shapes, but these are very tiny in visual examination. Notably, the FID score on the validation set does not display a marked reduction. Collectively, these results imply that although attention can change the local appearance/detail and training dynamics, the visual differences are otherwise minimal in the current set of experiments.

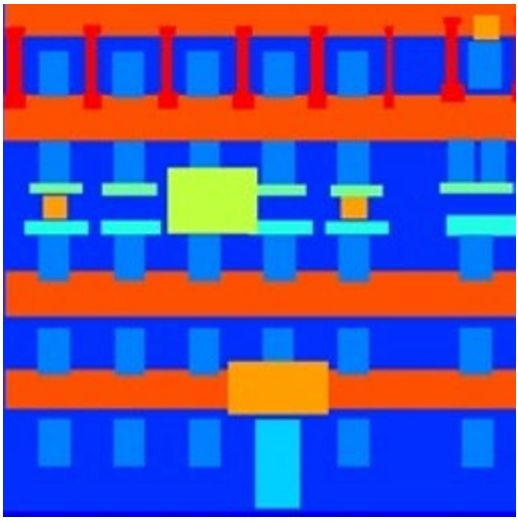

Fig. 1. The real_A images from the 200th training epoch of the Attention-Pix2Pix model (Picture credit: Original)

Fig. 2. The fake_B images from the 200th training epoch of the Attention-Pix2Pix model (Picture credit: Original)

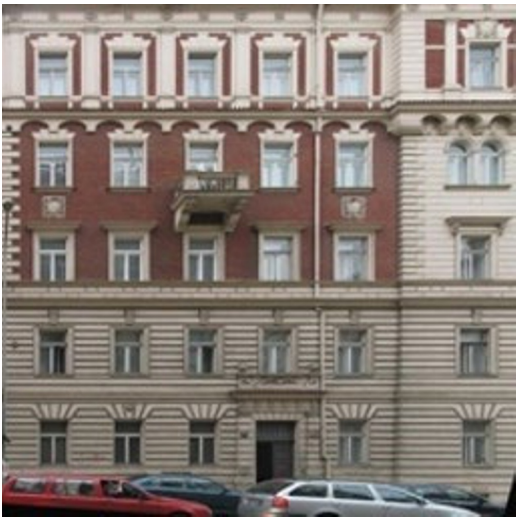

Fig. 3. The real_B images from the 200th training epoch of the Attention-Pix2Pix model (Picture credit: Original)

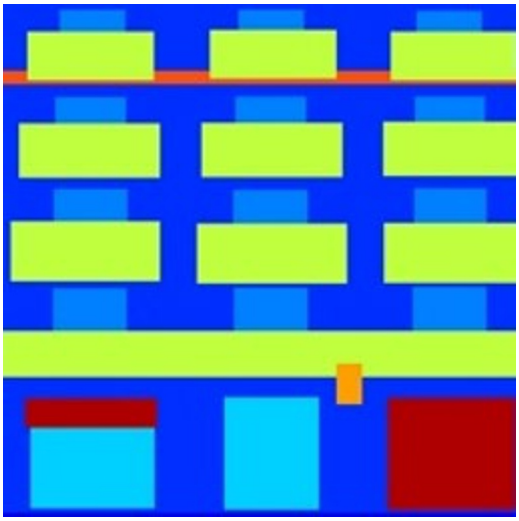

Fig. 4. The real_A images from the 200th training epoch of the Pix2Pix model (Picture credit: Original)

Fig. 5. The fake_B images from the 200th training epoch of the Pix2Pix model (Picture credit: Original)

Fig. 6. The real_B images from the 200th training epoch of the Pix2Pix model (Picture credit: Original)

4 Discussion

In this work, the paper proposes the incorporation of a self-attention module to the Pix2Pix model and discusses the diversity of the results by means of loss curves, synthesized images and the FID measure. The experimental results imply that attentions might cause some variation in the training dynamics. More concretely, the G_L1 loss becomes overall lower, the G_GAN curve appears more smooth, and the gap between D_real and D_fake tends to get more stable over the training, which may indicate a more balanced competition between the generator and the discriminator. Nevertheless, these modifications of training dynamics do not seem to bring significant gains in terms of perceptual output quality. The synthesized images look visually similar to the ones from baseline model and the FID score on val set does not show any significant drop.

Several factors may help explain this phenomenon. First, the Facades dataset has very structured label–image pairs, where the baseline Pix2Pix already achieves a strong result. In this setting, the long-range dependencies are less significant, and the effect of the attention may be limited. Second, the attention module is used only at the bottleneck of the U-Net generator: although this captures long-range dependencies at a very low resolution, its effect on high-resolution outputs is likely very minor. Thirdly, since the benefits brought by attention (i.e. better stability and local consistency) are not obviously related to the aspects to which the FID is most sensitive, small changes in the metric could be expected.

Overall, the results suggest that, for the present study, attention might have mainly rendered the optimization procedure more stable and local details more refined, rather

than bringing significant perceptual quality improvement on a structurally simple dataset. Future work will investigate extending the technique to more complex and varied datasets (e.g. Cityscapes [14]) to ascertain more general advantages, placing attention modules at other layers of the decoder to further enhance the modeling of high-resolution features, integrating attention with perceptual or feature-matching loss to enhance perceptual realism, and conducting a systematical ablation on different kinds of attention modules and insertion positions to gain further insights of their functionalities in conditional GANs.

5 Conclusion

The paper studies the beautification of urban building facades from a single image based on Pix2Pix, an image-to-image translation model. By analyzing training loss dynamics in detail (i.e. G_L1, G_GAN, D_real and D_fake) along with some qualitative comparisons of images and FID scores, a few interesting conclusions can be drawn.

First, it seems that introducing attention causes some teaching inconsistencies. The Attention-Pix2Pix model is also more competitive than baseline Pix2Pix model with the G_L1 loss gradually decreasing in a more steady manner, the graph of G_GAN appears smoother and the gap between D_real and D_fake becomes smaller, which may suggest a more balanced adversarial training process. The results may indicate that the attention module facilitates the model to learn global structure and long inter-dependencies which leads to a more stable optimization.

Second, although the training patterns exhibit some disparities, the quality of the output images remains relatively consistent and is comparable to the baseline in terms of perceptual quality. The enhanced attention model improves slightly in a few local areas, but the visual results are essentially consistent and the FID score does not reduce significantly. For example, this might imply that for a fairly structured data set such as Facades, the benefits of global context modeling might not immediately be reflected in perceptual gains.

In summary, under the experimental settings of this study and on the Facades dataset, the introduction of self-attention appears to influence training dynamics and local feature behavior, while its impact on overall perceptual image quality remains limited.

Future extensions of this work may explore evaluating the method on more complex or diverse translation tasks (e.g., semantic segmentation-to-photo on Cityscapes), incorporating multi-level or multi-head attention within the decoder, combining attention with perceptual or feature-matching losses to boost semantic fidelity, and performing ablation studies to systematically analyze the effect of different attention types and insertion locations. This study provides an empirical view of how attention mechanisms could affect conditional GAN training and offers insights for future architecture design and performance enhancement.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y.: Generative Adversarial Nets. In Advances in Neural Information Processing Systems (Vol. 27, pp. 2672-2680). Montréal, Canada: MIT Press. (2014)
2. Mirza, M., & Osindero, S.: Conditional Generative Adversarial Nets. arXiv:1411.1784. (2014)
3. Zhu, J. Y., Park, T., Isola, P., & Efros, A. A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2223-2232). Venice, Italy. (2017)
4. Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A.: Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1125-1134). Honolulu, Hawaii. (2017)
5. Ronneberger, O., Fischer, P., & Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) (Vol. 9351, pp. 234-241). Munich, Germany: Springer. (2015)
6. Radford, A., Metz, L., & Chintala, S.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the International Conference on Learning Representations (ICLR) (Workshop Track). San Juan, Puerto Rico. (2016)
7. Arjovsky, M., Chintala, S., & Bottou, L.: Wasserstein Generative Adversarial Networks. In Proceedings of the International Conference on Machine Learning (ICML) (Vol. 70, pp. 214-223). Sydney, Australia: PMLR. (2017)
8. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X.: Improved Techniques for Training GANs. In Advances in Neural Information Processing Systems (Vol. 29, pp. 2234-2242). Barcelona, Spain: MIT Press. (2016)
9. Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B.: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 8798-8807). Salt Lake City, Utah. (2018)
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I.: Attention Is All You Need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998-6008). Long Beach, California: MIT Press. (2017)
11. Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A.: Self-Attention Generative Adversarial Networks. In Proceedings of the International Conference on Machine Learning (ICML) (Vol. 97, pp. 7354-7363). Long Beach, California: PMLR. (2019)
12. Liu, C. C., Liao, Z., & Efros, A. A.: Deep Learning for Facade Parsing and Understanding. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 320-329). Venice, Italy. (2017)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6626-6637). Long Beach, California: MIT Press. (2017)
14. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, H. R., & Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3213-3223). Las Vegas, Nevada. (2016)