UDC 67

A NEW PARADIGM FROM SKETCH TO REALISTIC IMAGES WITH TRANSFORMER

Yifeng Sun，Machinery 2263，Research 1
Zhiyong Zhou，Postdoctoral，VicePresident
Shanghai Dianji University
Shanghai，China

Abstract： To overcome the challenging problem of synthesising real images from sketches, we have proposed a new deep learning network, VGAN, in which we introduce a new module, Vit-Conv Block, into the GAN network. The attention mechanism is added to the Vit-Conv Block, allowing the module to better extract features from sketches. The new network based on the VGAN module is able to handle the problem from sketch to image very well. The final result is that the VGAN model produces images with higher quality and smoother edge processing than pix2pix.

Introduction： How do we draw a picture quickly when only sketches are available. To many novices this seems unachievable. But in recent years neural networks have provided a reasonable approach to this idea, with generative adversarial networks [1] showing great potential. is "adversarial training is the coolest thing since sliced bread". It uses two neural networks, one of which (we call Generator) is pitted against another (Discriminator), and with GAN we can generate realistic images.

We present a new network that improves on the GAN model by fusing the Transformer [2] with the CNN. The aim is to make the images generated by our VGAN model as close as possible to the real samples.

In this paper, we propose a new architecture combining Transformer as well as CNNs, and we summarize the main contributions of this paper as follows:

1. 1. A new feature extraction module, Vit-Conv, has been built to better extract image features.

2. The new GAN network we have built can generate more realistic images.

Related work. Image to Image Translation: Sketch to image aims to learn the mapping and elimination of domain gaps between hand drawn and real images, and is typically modeled as an image to image conversion task. Use GANs for image to image conversion. Generative adversarial networks have shown great potential in generating natural and realistic images. The pix2pix [3] work demonstrates a simple

method of converting one image into another using conditional GANs. SketchyGAN [4] is a pioneering work dedicated to solving sketch input translation. However, the training process requires a dataset composed of various paired sketch image data, which is difficult to collect for the training dataset.

Methods: 1.Model Architecture: In this article, we propose a new architecture of GAN model combined with Transformer - VGAN. We introduced Transformer in the Generator and adopted the Vit architecture. Firstly, the input image will go through two Vit Conv Blocks [5] to obtain global features from shallow layers, preserving more spatial information. Then the Unet [6] architecture is used to extract rich Semantic information. Finally, the information extracted from Vit Conv Block and DeConv Block is feature spliced with UpConv Block through up sampling and jump connection to enrich global features and spatial information as much as possible. Discriminator adopts the classic Unet architecture. The overall architecture of VGAN will be shown in Figure 1. A more detailed model structure will be explained below.

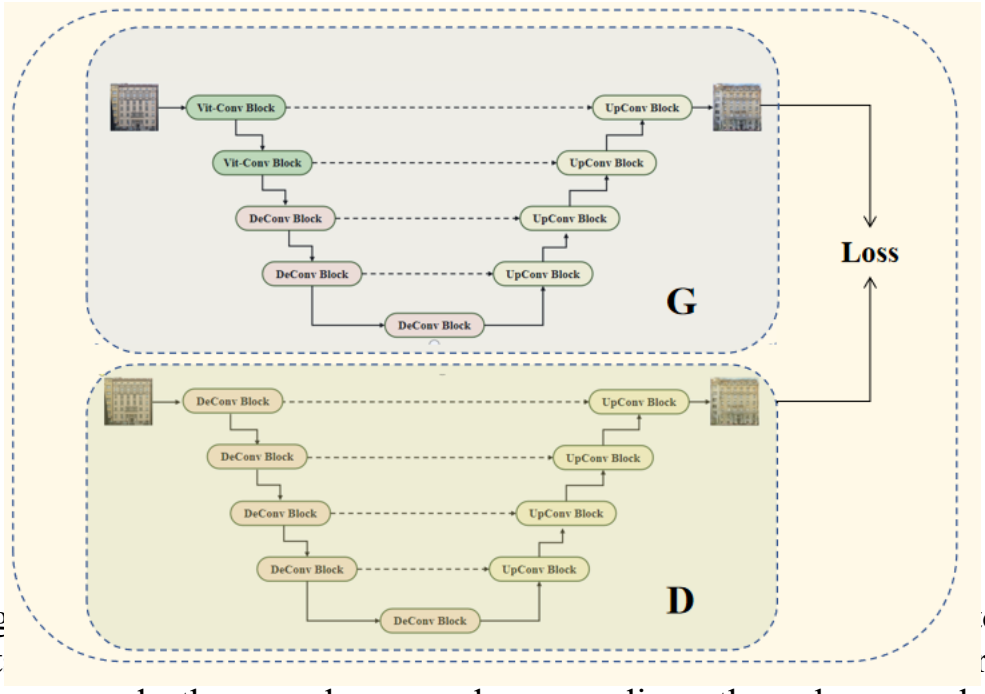


Fig or, and D represent rough the Transformer, and then undergoes downsampling through convolution. In Discriminator, we adopt a pure convolutional composition.

2.Vit-Conv Block

The Transformer application in machine vision was proposed by the Google team [6]. We have made certain adjustments to the structure of Vit based on our data. Our structure is shown in the figure.

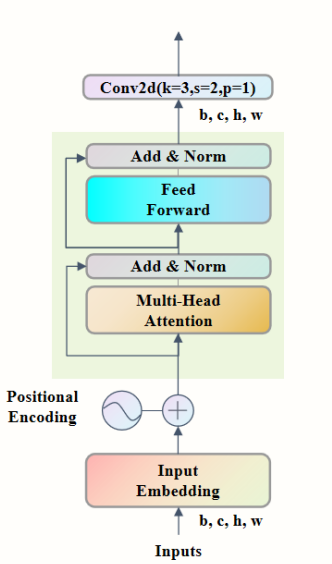


Figure 2. According to the process described above, the Vit Conv Block can be divided into the following steps:

1. Patch embedding: The input image size is 256 × two hundred and fifty-six × 3. Divide the image into 16 × A fixed size patch of 16, which means the length of the input sequence is 256, and the dimension of each patch is 16x16x3=768. The dimension of the linear projection layer is 768xN (N=768), so the input dimension after passing through the linear projection layer is still 256x768, which means there are a total of 256 tokens, each with a dimension of 768. In addition, a special character cls needs to be added, resulting in a final dimension of 257x768. The purpose of doing so is to transform visual problems into seq2seq problems.

2. Position encoding: It is also important for Vit position encoding. Position encoding can be understood as a data list of N rows, with the size of N consistent with the input length. After adding position encoding, the dimension remains unchanged and remains 257 × 768 (The operation of position encoding is sum).

3. Multi head attention: The multi head attention mechanism first maps the input vectors to q, k, and v. We set 12 heads, and the dimension of qkv is 257 × (768/12), i.e. 257 × 64, a total of 12 groups.

4. MLP: The function of MLP is to scale the dimensions, allowing for sufficient crossing between different dimensions of feature vectors, and reducing the dimensionality method to 257 × 3072, then restore the dimension to 257 × 768, this is to enable the model to capture more information on nonlinear and composite features.

5. Conv2d: Convolutional kernel size 3 × 3. The stripe is 2, and the padding is 1. After Conv2d, the size of the image can be compressed to half its original size and the number of channels can be changed.

3.DeConv Block

The structure of Conv2d is a combination of point multiplication, addition

and nonlinear activation function between a two-dimensional convolution kernel and input data. Simply put, Conv2d can fuse, reduce dimensionality, and extract feature information from the pixel information of input data through convolutional kernel operations. This process involves multiple layers of convolution and pooling operations, ultimately resulting in a highly abstract feature representation for tasks such as classification, detection, or segmentation. The DeConv Block consists of two Conv2d structures, as shown in Figure 2.

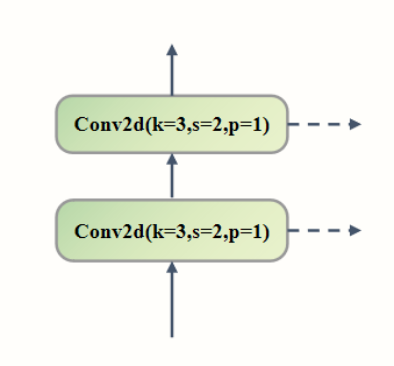


Figure 3. DeConv Block: In the DeConv Block module, we used two convolutions, with a convolution kernel size of 3 × 3. The stripe is 2, and the padding is 1. After each convolution, it will jump connect to the UpConv Block.

4.UpConv Block

ConvTranpose2d is a two-dimensional transposed convolution operation, also known as deconvolution. It is widely used in deep learning for signal reconstruction and generation tasks such as image and speech. Unlike ordinary convolutions, transposed convolutions expand the size of the input tensor and are typically used for upsampling operations.

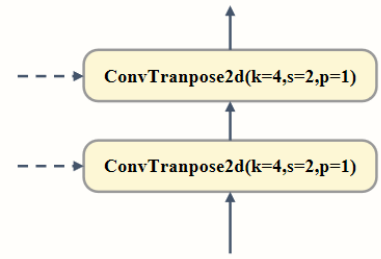


Figure 4. UpConv Block: In the UpConv Block module, we used transposed convolution twice, with a convolution kernel size of 3 × 3. The stripe is 2, and the padding is 1. After each transposed convolution, a skip connection is made to the DeConv Block.

Experiment

1.Datasets：Facades dataset consists of 506 Building Facades & corresponding Segmentations with split into train and test subsets(train sets:400,test sets:106).

Table 1. Partition of Datasets

	all	train	test
number	506	400	106

2.Training setup: The training model used two GTX 2080Ti blocks, cuda 11.6, pytorch 1.13.1, and Linux 4.15.0, which are divided into generator loss and discriminator loss in our model. Generator losses are divided into adversarial losses and pixel losses as follows:

$$L_{adv} = -E_{A \in p_{data}(A)}[\log D(G(A), A)] \quad (1)$$

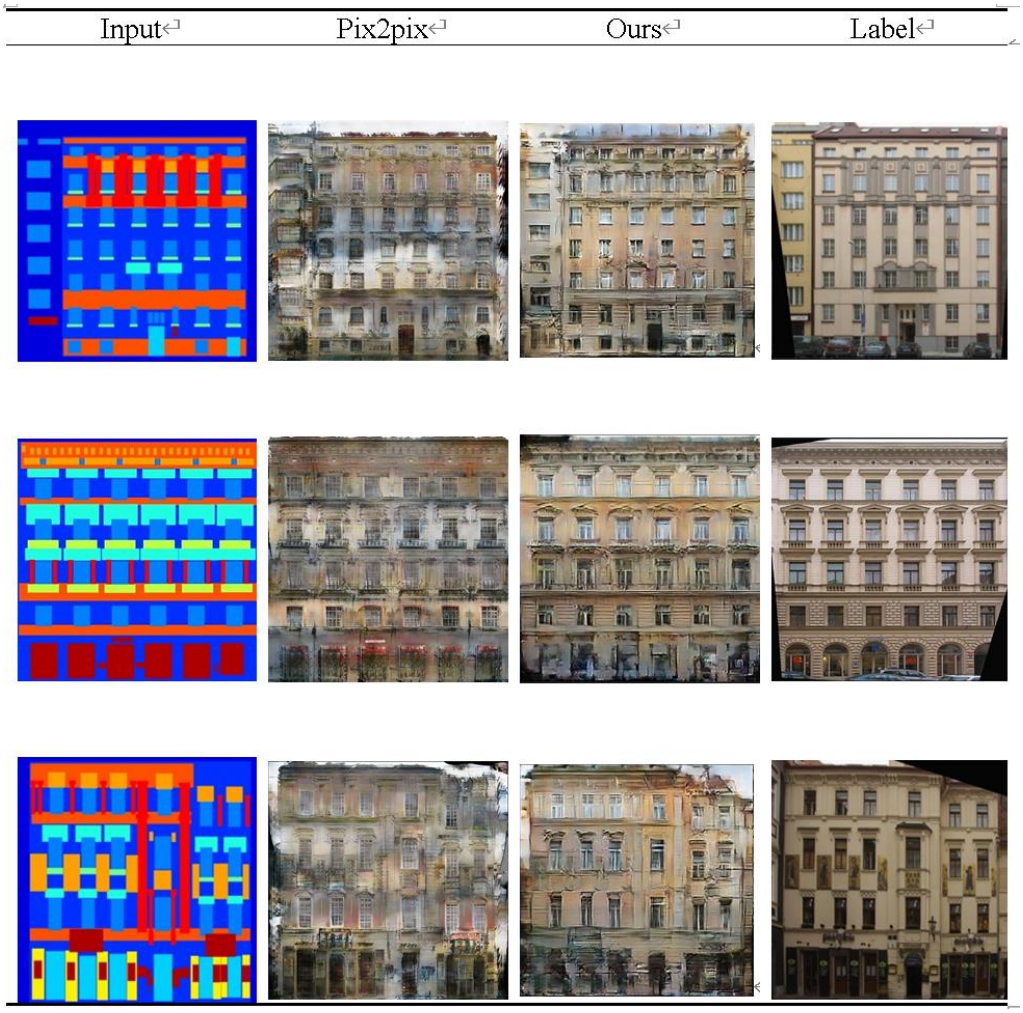
$$L_{pix} = -E_{A \in p_{data}(A), B \in p_{data}(B)}[\| B - G(A) \|_1] \quad (2)$$

The total loss of the generator is:

$$L_G = L_{adv} + \lambda * L_{pix} \quad (3)$$

Result. Here we demonstrate the generation results of pix2pix and our model on the Facades dataset (in order to compare the fairness of the experiment, we adopted a completely consistent training plan). Obviously, under the same training parameters, our model generates more delicate images, smoother edge processing, and less noise. The generation result of pix2pix is blurry and the edges are messy. This indicates that our model architecture is effective.

Table 2. Results display of Pix2pix and Ours on the Facades dataset



Conclusion. We propose a GAN network framework that combines vit, with both the generator and discriminator adopting a U-shaped structure. Our Vit Conv Block can extract the global context information, combined with the convolution operation, to provide more abundant Semantic information for the generator in the model, which is more accurate than the features extracted only through convolution. Therefore, our method achieved convincing results on the Facades dataset. On the basis of this work, there are many directions that can be expanded in the future. For example, generating higher resolution images and incorporating textual information. In addition, improving the reliability of GAN network training is also a worthwhile research topic.

References

1. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
2. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in

neural information processing systems, 2017, 30.
3. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Ad- versarial Networks. In CVPR, 2017. 1, 2, 3, 4, 6, 7, 19
4. Chen W, Hays J. Sketchygan: Towards diverse and realistic sketch to image synthesis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9416-9425.
5. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
6. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020