# **A Hybrid Model Integrating Vision Transformers and Convolutional Neural Networks for Text-to-Image Synthesis**

## **Abstract**

Transformer models have received a lot of attention from computer vision researchers recently because of its strong global feature extraction ability for global features based on large amounts of data pre-training and has been widely used for several traditional problems related to Convolutional Neural Networks (CNNs) have a receptive field focused on local context while transformer have receptive field concerned on the global context. At the same time, there has also been a vast increase in the capabilities of generating images based on GANs.

This paper introduces a new text-to-image synthesis model that combines a transformer-based generator and an improved CNN-ViT based hybrid discriminator. The generator uses multi-layer transformers to extract contextual dependencies from input texts, and the improved discriminator is a convolutional layer fused with Vision Transformer (ViT). Through combining both local and global information, the discriminator enhances its performance in distinguishing between the real and fake images.

**1. Introduction**

Generative models based on deep learning algorithms have been recently widely explored methods in the field of text-to-image synthesis.The traditional Convolutional Neural Networks (CNNs) have provided significant results in generative tasks though they have a drawback of local receptive fields. In contrast, the Vision Transformers (ViTs) are capable of learning global dependencies that exist across the entire image, but often, they all need a large amount of data and a substantial amount of computational power. Transformers which are often stacked with Self-attention modules rarely use convolution operations. Thus, this model shares well-suited characteristics of both paradigms: the local sensitivity of CNNs and the global awareness of the transformers.

Our work extends the latest advancements in the area of hybrid architectures in which both these models are incorporated across the generative and discriminative components. In particular, we suggest an improved discriminator that uses the image encoder built on convolutional layers sequentially after the Vision Transformer module that accounts not only for the spatial features but also for the contextual ones. Combining this with a transformer-based generator, our model outperforms its counterparts in transcribing text to realistic images while proving stable during its training.

**2. Related Work**

The access to large-scale as well as high-quality labelled data has made way for the emergence of several fields such as image classification. Earlier, CNNs’ performance was slow on the visual tasks because of the restricted power of GPUs and the available storage space. However, it was not until Krizhevsky et al. [[1](#_ufu3peyq2mb3)] who demonstrated that convolutions were not only useful for text data processing but also for supervising the learning process through gradient backpropagation. This led to the extensive use of CNNs and their derivatives as the baseline network for computer vision. However, it is not perfect and is replaced by the transformer when it comes as a solution.

**2.1 Convolution Neural Networks**

Advances in computer hardware have seen CNNs scale in depth and width and, therefore, post massive improvements in their performance. VGG [[2]](#_ufu3peyq2mb3) proved that to deepen the network, instead of using large-scale convolutions, stacked 3 × 3 convolutions can enhance the performance of deep networks in image feature learning. Following this idea, GoogLeNet [[3]](#_ufu3peyq2mb3) built up multi-branch paths via convolution kernels of various sizes to enhance recognizing speed. CNNs were developed based on multi-branch deep networks which were manually designed and offered another way to enhance the network capacity to cover diverse scenes. Based on the above, Szegedy et al. put forward InceptionV2, V3, V4 [[4–6]](#_ufu3peyq2mb3) to construct efficient networks. Nevertheless, this idea of networks being accurate decreases and evens out as the networks unsheathe a deeper appearance. As for deeper CNNs, overfitting is not the only problem, the network also experiences gradients disappearing or exploding. To address such issues, He et al. [[7]](#_ufu3peyq2mb3) proposed a residual learning network – ResNet – ResNet is able to directly connect to hundreds of layers of networks through residual connection, defined as the identity residual function linking it to the received information. ResNetX [[8]](#_ufu3peyq2mb3) went a step further than this by splitting the residual convolution submodule into groups, in this case identical, and then merging them as shown below to enhance the expression capability of the model.DenseNet [[9]](#_ufu3peyq2mb3), on the other hand extended feature reuse concept in feedforward manner by connecting each convolution layer with the previous output features. More work has also been done involving CNNs further based on these revolutionary ideas but here it is about manually designing different models for tasks. For instance, RepVgg [[10]](#_ufu3peyq2mb3), ConvNet [[11]](#_ufu3peyq2mb3), and EfficientNet [[12]](#_ufu3peyq2mb3) are examples of such models.

When observing an image, humans ignore the rest of the background and concentrate on areas of interest. In order to implement such attention mechanisms to the network to focus on the salient features of the feature space and channel, the opportunity of focusing more on the features of the sample should be provided to the network. Hu et al. [[13]](#_ufu3peyq2mb3) put forward SeNet, which relied on both compression and excitation blocks to construct inter-channel correlation appropriately and, thereby, highlight some details of information features. Wang et al. [[14]](#_ufu3peyq2mb3) introduced Efficient Channel Attention for Deep CNNs (ECA-Net) to enhance channel attention. ECA-Net employs a cross-channel interaction approach, which features learnable one-dimensional convolution kernel size. CBAM [[15]](#_ufu3peyq2mb3) used intermediate feature maps between the channel and spatial order, multiplied its weighted by the input to eliminate the features components not needed. Recently, Li et al. [[16]](#_ufu3peyq2mb3) applied dynamic convolution kernels with various receptive fields to choose attention feature maps and combine multi-scale information with linear operation. SegNext [[17]](#_ufu3peyq2mb3) also designed an efficient convolutional network by incorporating multi-scale convolutional features to elicit spatial attention. Zhu et al. [[18]](#_ufu3peyq2mb3) proposed AW-convolution to solve the issue related to limited capacity of attention maps. However, the attention mechanism weights must be concatenated with these related convolutions and are supplementary to the CNN architecture.

**2.2 Vision Transformers**

Compared with convolution-based attention structures, the self-attention structure of the transformers is a separate feature extractor. Transformers were introduced in a paper for the first time by Vanswani et al. [[19]](#_ufu3peyq2mb3), and these are more popular in NLP that includes BERT [[20]](#_ufu3peyq2mb3) and others such as GPT-3 [[21]](#_ufu3peyq2mb3). Transformers should be trained on a vast database before being tuned on a certain task dataset, where the model parameters are moved. Dosovtskiv proposed ViT, which applies the decoder rules of the transformer to image processing by dividing the image into non-overlapping image patch sequences and using self-attention maps to analyse the correlation among the sequences. Most transformers require pre-training on large datasets which consumes a lot of resources and time. In related fields, it is difficult to obtain a large amount of high-quality labelled images due to factors such as cost, privacy, security, or ethical concerns. To address these challenges, researchers have explored combining transformers with mature CNNs. Models such as CeiT [[22](#_ufu3peyq2mb3)], BoTNet [[23](#_ufu3peyq2mb3)], CMT [[24](#_ufu3peyq2mb3)], and Next-ViT [[25](#_ufu3peyq2mb3)] attempted to combine transformers with CNNs, using convolutions to extract low-level and local features with considerable accuracy. As a result, combining transformers and CNNs is a promising research direction.

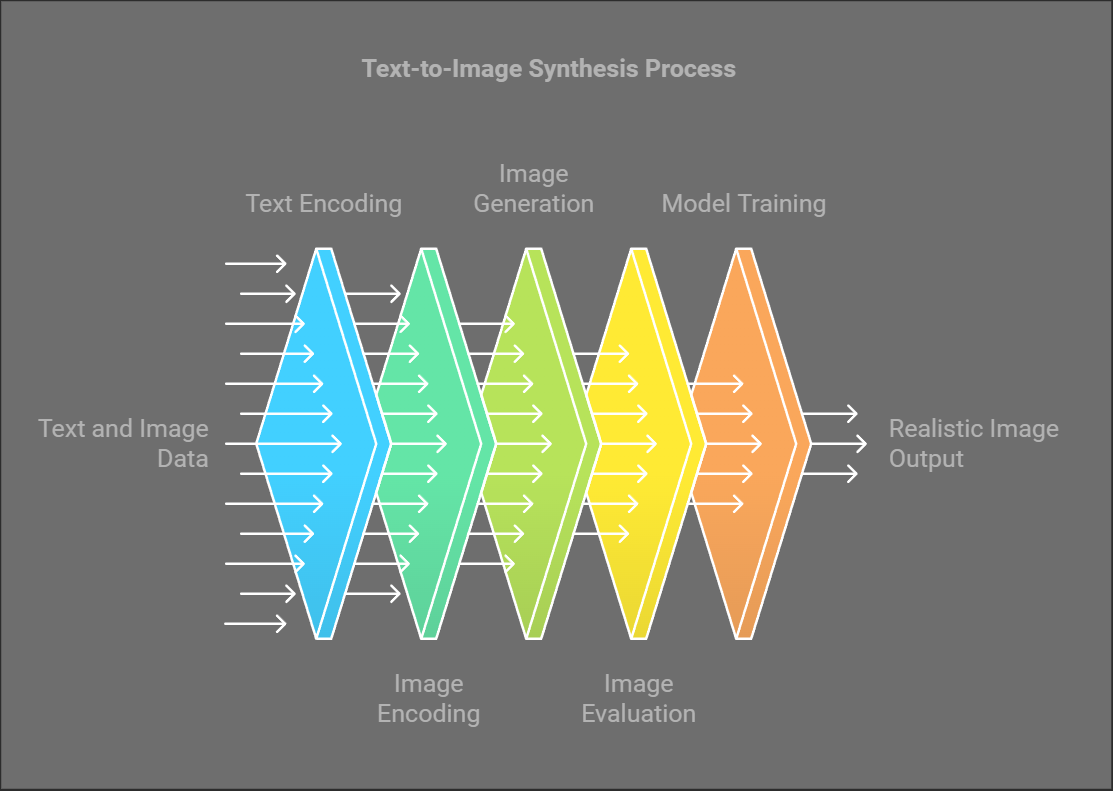


Fig-1: Basic Workflow of the model

**2.3 Combining CNNs and Transformers**

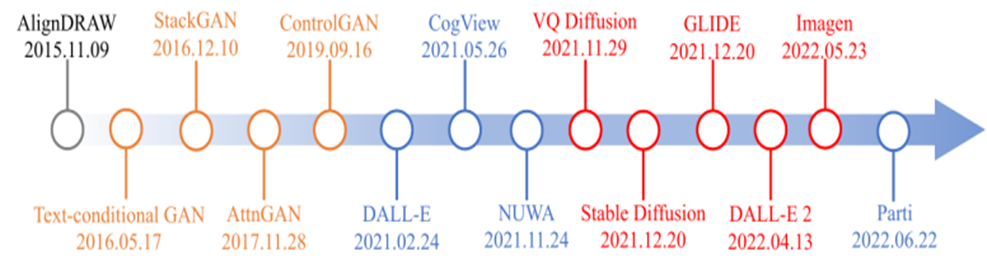
Several studies have explored the integration of CNNs and transformers to address various computer vision tasks. For example, the increased transformation operation in CoAtNet [[26]](#_ufu3peyq2mb3) and CvT [[27]](#_ufu3peyq2mb3) shows that hybrid models offer even better efficiency and generalisation. Recent works have also explored the usage of discriminators that are based on CNN, together with the generators that rely on the transformer in GANs for image synthesis as described in the paper by Combining Transformer Generators with Convolutional Discriminators [[28]](#_ufu3peyq2mb3). However, these models often struggle with training instability or fail to fully leverage global dependencies during discrimination. Our work extends these approaches by employing a hybrid discriminator with enhanced feature extraction capabilities, resulting in better stability and synthesis quality. 

Fig.2 Text-to-image generative models

## **3. Proposed Work**

### **3.1 Overview**

Our proposed hybrid architecture integrates a transformer-based generator with an enhanced discriminator featuring a hybrid image encoder. This dual-architecture design addresses the key limitations of prior models by ensuring that both the generator and discriminator exploit the complementary strengths of CNNs and transformers.

### **3.2 Transformer-Based Generator**

### The generator employs a BERT-based text encoder to extract meaningful embeddings from the input text. These embeddings are combined with a noise vector to initialize the generative process. The core of the generator consists of multi-layer perceptrons (MLPs) and transformer layers, which allow the model to capture long-range dependencies between text features and visual content. The final output is reshaped into image data and scaled using a tanh activation function to ensure pixel values are in the range of [-1, 1].

### **3.3 Enhanced Discriminator with Hybrid Image Encoder**

The discriminator plays a critical role in ensuring the quality of generated images by distinguishing between real and synthetic data. To enhance its performance, we employ a hybrid image encoder within the discriminator. This encoder first processes the input image using convolutional layers, capturing fine-grained spatial patterns. The feature map is then passed through a ViT module, which extracts global contextual features, allowing the discriminator to make more informed decisions.

The enhanced discriminator combines these local and global insights to evaluate the authenticity of images with higher accuracy. By concatenating the image features with text embeddings, the discriminator also ensures that the generated images align semantically with the input text. This design enables the discriminator to act as a more reliable judge during adversarial training.

**3.4 Loss Function**

We are using Binary Cross-Entropy (BCE) loss for training the discriminator and generator in the GAN setup. The BCE loss measures the difference between the predicted probabilities (real or fake) and the actual labels.

We can describe BCE loss as:

BCE(y,ŷ) = - [y ​log(ŷ​) + (1− y) log(1− ŷ ​)]

### **4. Experimental Results**

To evaluate the proposed hybrid architecture, we conducted experiments on the CIFAR-10 dataset, focusing specifically on generating images of the "cat" class. This choice allows us to assess the model's ability to capture fine-grained visual details and ensure semantic alignment with the given textual descriptions. The images were generated at a resolution of 64x64 pixels, which balances computational efficiency with visual clarity.

#### **4.1 Experimental Setup**

The training process utilized the following configuration:

* **Data Preprocessing**: The CIFAR-10 dataset was filtered to retain only images of the "cat" class. Images were resized to 64x64 pixels and normalized to enhance training stability.
* **Model Training**: The model was trained for 100 epochs using a batch size of 16. The learning rates for the generator and discriminator were set to 0.0001 with Adam optimization (betas = (0.5, 0.999)). Binary Cross-Entropy (BCE) adversarial loss was used to optimise both the generator and discriminator.
* **Evaluation Metrics**: Performance was assessed based on the visual quality of the generated images, the alignment with textual descriptions, and the convergence speed of the adversarial losses (d\_loss and g\_loss). We also performed a qualitative analysis to visually inspect the generated images

Since we are using a hybrid approach with a Vision Transformer (ViT) and convolutional layers, we can describe the image encoding as:

Encoded Features = FC**(**ViT**(**Conv**(** 𝑋 **)))**

where 𝑋 is the input image, Conv represents the convolutional layers, ViT refers to the Vision Transformer, and FC is the final fully connected layer for feature compression.

#### **4.2 Image Quality and Semantic Alignment**

The transformer-based generator demonstrated a superior ability to produce realistic images that corresponded accurately with the input text descriptions. Compared to traditional CNN-based GANs, our generator was able to capture nuanced textures and details, such as the shapes and fur patterns of cats, making the generated images appear more lifelike. This improvement is attributed to the generator's use of transformer layers, which model the relationships between the text features and visual content more effectively than convolutional approaches alone.

#### **4.3 Enhanced Discriminator Performance**

The enhanced discriminator featuring a hybrid CNN-ViT encoder showed significant improvements in distinguishing between real and synthetic images. By combining local feature extraction through convolutional layers with global context modelling using the Vision Transformer, the discriminator was able to better identify subtle artefacts in the generated images. This resulted in:

* **Improved Stability**: The adversarial training process displayed more stable convergence, with the discriminator and generator losses oscillating less compared to traditional GAN setups. The incorporation of the ViT in the discriminator enabled it to detect complex patterns across the entire image, leading to more reliable feedback during training.
* **Faster Convergence**: The model achieved noticeable improvements in image quality within the first 30 epochs, where traditional GAN models often require more training epochs to produce visually comparable results. The combination of CNNs for local features and transformers for global understanding allowed the discriminator to provide more meaningful gradients to the generator, accelerating the learning process.

**5. Results**

In this section, we present the experimental findings of our hybrid architecture on the CIFAR-10 dataset, focusing specifically on the "cat" class to validate its performance in generating realistic images from text descriptions.   
  
**5.1 Training Dynamics**

The training process was monitored over 100 epochs. The generator and discriminator losses were tracked to ensure stable training and convergence. Figure 3 displays the training loss curves for the generator (g\_loss) and discriminator (d\_loss), respectively. The enhanced discriminator with a hybrid image encoder facilitated more stable adversarial training, as shown by the reduced oscillations in loss values compared to models without transformer components.

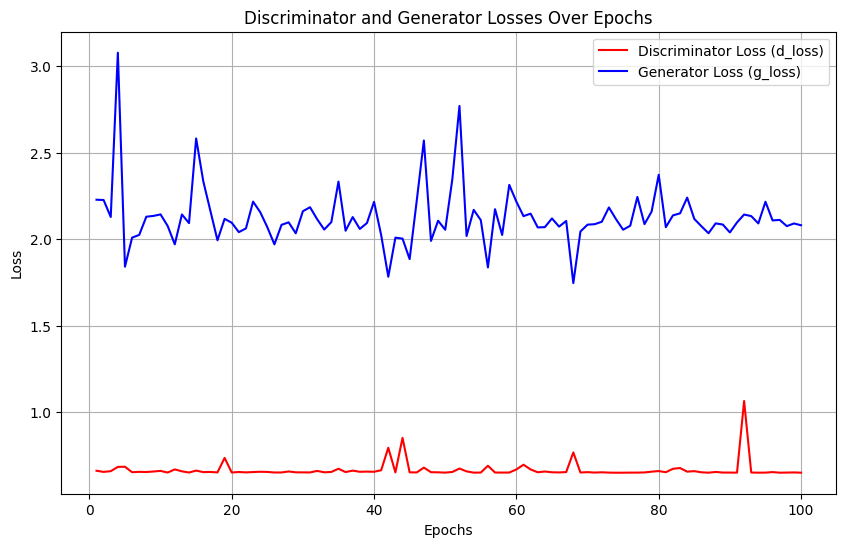


Fig.3 Discriminator and Generator losses

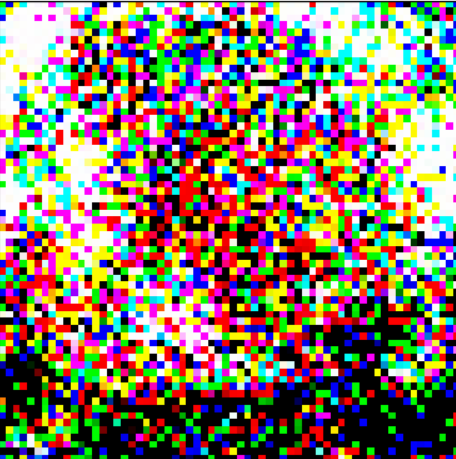


Fig. 4 Generated image at 1 epoch Fig.5 Generated image at epoch 100

## **6. Conclusion**

This paper presents a novel hybrid architecture that integrates a transformer-based generator with an enhanced CNN-ViT hybrid discriminator for text-to-image synthesis. Our approach leverages the complementary strengths of CNNs and transformers, resulting in superior image quality and stable adversarial training. The enhanced discriminator, with its ability to capture both local and global features, plays a crucial role in improving synthesis quality. Future work could explore scaling this model to higher resolutions and applying it to more diverse datasets. Our findings demonstrate that the hybrid approach offers a promising direction for advancing the state of the art in text-to-image generation.

## **References**

1. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 6(60):84–90. <https://doi.org/10.1145/3065386>
2. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Proceedings of international conference on learning representations (ICLR)
3. Szegedy C, Liu W, Jia Y, et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1-9. [https://doi.org/10.1109/CVPR. 2015.7298594](https://ieeexplore.ieee.org/document/7298594)
4. Ioffe S, Szegedy C (2015) Batch normalisation: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the international conference on machine learning (ICML), pp 448–456
5. Szegedy C, Vanhoucke V, Ioffe S, et al (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2818-2826. <https://doi.org/10.1109/CVPR.2016.308>
6. Szegedy C, Ioffe S, Vanhoucke V, et al (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence, 31(1)
7. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 770-778. [https://doi.org/ 10.1109/CVPR.2016.90](https://ieeexplore.ieee.org/document/7780459)
8. Xie S, Girshick R, Dollár P, et al (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 5987-5995. <https://doi.org/10.1109/CVPR.2017.634>
9. Huang G, Liu Z, VanDerMaaten L, et al (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2261-2269. https:// [doi.org/10.1109/CVPR.2017.243](http://doi.org/10.1109/CVPR.2017.243)
10. Ding X, Zhang X, Ma N,et al (2021) Repvgg: Making vgg-style convnets great again. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 13728-13737. <https://arxiv.org/abs/2101.03697>
11. Liu Z, Mao H, Wu CY, et al (2022) A convnet for the 2020s. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 11966-11976. [https://doi.org/10.1109/CVPR52688. 2022.01167](https://ieeexplore.ieee.org/document/9879745)
12. Tan M, Le Q (2019) Efficient Net: Rethinking model scaling for convolutional neural networks. In: Proceedings of the international conference on machine learning (ICML), pp 6105–6114
13. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7132-7141.<https://arxiv.org/abs/1709.01507>
14. Wang Q, Wu B, Zhu P, et al (2020) Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 11531-11539. <https://doi.org/10.1109/CVPR42600.2020.01155>
15. WooS, Park J, Lee JY, et al (2018) Cbam: Convolutional block attention module. In: Proceedings of the european conference on computer vision (ECCV), pp 3–19
16. Li X, Wang W, Hu X, et al (2019) Selective kernel networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 510-519.<https://ieeexplore.ieee.org/document/8954149>
17. Guo MH, Lu CZ, Hou Q, et al (2022) Segnext: Rethinking convolutional attention design for semantic segmentation. Preprint at [arXiv:2209.08575](https://arxiv.org/abs/2209.08575)
18. Zhu B, Hofstee P, Lee J, et al (2021) An attention module for convolutional neural networks. In: Proceedings of 30th international conference on artificial neural networks, Bratislava, Slovakia, pp 167-168. <https://doi.org/10.1007/978-3-030-86362-3_14>
19. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Adv Neural Inf Process Syst (NeurIPS) 30
20. Devlin J, Chang M W, Lee K, et al (2018) Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP)
21. Brown T, Mann B, Ryder N et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst (NeurIPS) 33:1877–1901
22. Yuan K, Guo S, Liu Z, et al (2021) Incorporating convolution designs into visual transformers. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 559-568. <https://arxiv.org/abs/2103.11816>
23. Srinivas A, Lin T Y, Parmar N, et al (2021) Bottleneck Transformers For Visual Recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 16514-16524. https:// [doi.org/10.1109/CVPR46437.2021.01625](http://doi.org/10.1109/CVPR46437.2021.01625)
24. Guo J, Han K, Wu H, et al (2022) Cmt: Convolutional neural networks meet vision transformers. In: Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR),pp12165-12175. <https://doi.org/10.1109/CVPR52688.2022.01186>
25. Li J, Xia X, Li W, et al (2022) Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. Preprint at arXiv:2207.05501v4
26. Zihang Dai, Hanxiao Liu, Quoc V. Le, Mingxing TanCoAtNet: Marrying Convolution and Attention for All Data Sizes. [[2106.04803] CoAtNet: Marrying Convolution and Attention for All Data Sizes (arxiv.org)](https://arxiv.org/abs/2106.04803)
27. H. Wu et al., "CvT: Introducing Convolutions to Vision Transformers," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, [CvT: Introducing Convolutions to Vision Transformers | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/9710031)
28. Ricard Durall, Stanislav Frolov, Andreas Dengel, Janis Keuper [[2105.10189v1] Combining Transformer Generators with Convolutional Discriminators (arxiv.org)](https://arxiv.org/abs/2105.10189v1)