Recent Advances in Vision Transformer: A Survey and Outlook of Recent Work

Khawar Islam

Abstract

Vision Transformers (ViTs) are becoming more popular and dominating technique for various vision tasks, compare to Convolutional Neural Networks (CNNs). As a demanding technique in computer vision, ViTs have been successfully solved various vision problems while focusing on long-range relationships. In this paper, we begin by introducing the fundamental concepts and background of the self-attention mechanism. Next, we provide a comprehensive overview of recent top-performing ViT methods describing in terms of strength and weakness, computational cost as well as training and testing dataset. We thoroughly compare the performance of various ViT algorithms and most representative CNN methods on popular benchmark datasets. Finally, we explore some limitations with insightful observations and provide further research direction. The project page along with the collections of papers are available at https://github.com/khawar-islam/ViT-Survey

Keywords: Vision transformer, self-attention, convolutions, long-range relationship.

---

1. Introduction

Transformer [1] was introduced in the field of NLP on translating English-to-German language task. This transformer is solely based on attention mechanism, instead of convolution layers. It comprises of an encoder and decoder modules. Both modules are consists of attention layers with feed-forward network. Before, inserting into a transformer, every word in sentence embedded into vector $d_{model} = 512$ dim. Recently, [2] applied existing architecture of transformer

on computer vision task and achieved top performing results. Nowadays, ViTs have emerged and gave more promising results, can outperforms CNN. From an architectural point of view, the backbone of ViTs entirely made of self-attention mechanism and achieved remarkable success in visual tasks. There are two major concepts that significantly contributed towards the evolution of ViT models. (a)Self-attention mechanism where ViTs captured long-range of token dependencies in a global context, same as traditional recurrent neural networks. (b) The second important benefit is to train on large-scale unlabelled datasets and fine-tune on other tasks with small datasets.

2. Overview of Vision Transformer

Graphical overview of additional ViT steps are presented in Figure. 1. It requires several additional steps to utilize NLP transformer for computer vision tasks. The overall architecture of ViT can be presented into five main points:

- Split image into non-overlap/overlap patches (16×16, 32×32, etc.)

- Flatten patches and generate lower-dimensional linear embeddings from the flattened patches referred to (Patch Embedding).

- Add positional embedding and class token.

- Feed the sequence of patches into the transformer layer and obtain the output (label) through a class token.

- Pass class token values to MLP head to obtain final output prediction.

2.1. Patch Generation.

Consider an input image of size 112×112, we firstly generate 16×16 non-overlap/overlap patches. Thus, we can generate 49 patches and straightforward feed into the linear projection layer. Remember that, the number of color channels in each patch is 3. To obtain a long vector representation of each patch, the patches is fed into the projection layer and these representations have been presented in the Figure. 1 as
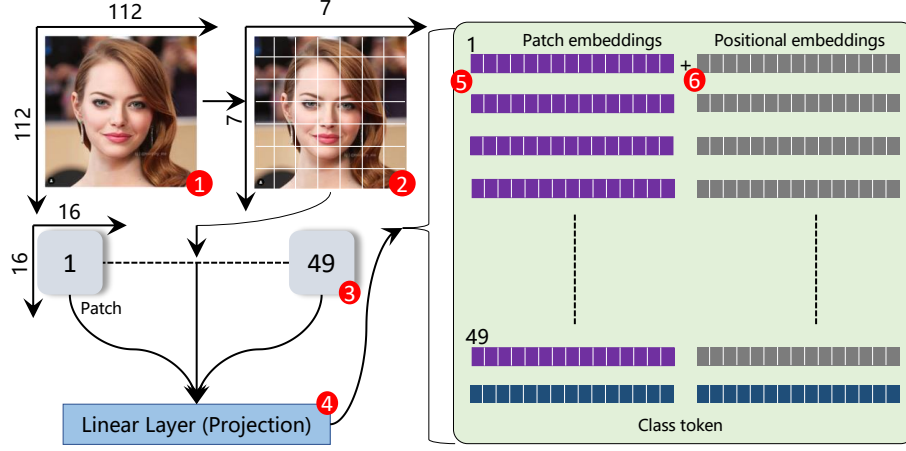
Figure 1: Additional steps of ViT.

2.2. Patch Embedding

The total number of overlap/non-overlap patches are 49, patch size with a number of channels is $16 \times 16 \times 3$. The size of the long vector of each patch is 768. Overall, patch embedding matrix is 49x196. Further, the class tokens have been added to the sequence of embedded patches and also added position Embedding. Without positional encoding, the transformer cannot retain the information and accuracy will be decreased to approx. 3%. Now, the patch embeddings size becomes 50 due to the additional class token. Finally, the patch embeddings with positional encoding and class token feed into the transformer layer and obtain the learned representations of the class token. Therefore, output from the transformer encoder layer is 1x768 which is passed to the MLP block to obtain the final prediction.

2.3. Transformer Encoder Layer.

Especially in ViTs, the most important component is the transformer encoder that contains MHSA and MLP block. The encoder layer receives combined embeddings (patch embeddings, positional embeddings, and class tokens) of shape 50 (49 patches and 1 [cls] token)$\times 768(16 \times 16 \times 3)$ as input. For all layers, the inputs and output of matrix shape 50x768 from the previous layer. In

3

Figure 2: Overview of transformer encoder block in vision transformer along with multi-head self-attention module.

ViT-Base architecture, there are 12 heads (also known as layers). Before feeding input into the MHA block, the input is being normalized through the normalization layer in Fig. 2. In MHA, the inputs are converted into $50 \times 2304(768 \times 3)$ shape using a Linear layer to obtain the Query, Key, and Value matrix. Then, reshaped these three QKV matrices into $50 \times 3 \times 768$ where each of the matrices of shape 50×768 presents the Query, Key, and Value matrices. Further, these Query, K, and Value matrices are reshaped again to $12 \times 50 \times 64$. Once, the Query, Key, and Value matrices are achieved, the attention operation inside the MHA module is performed using the below equation:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q} \cdot \mathbf{K}^\top}{\sqrt{d_k}}) \cdot \mathbf{V}, \tag{1}$$

Once we obtain the outputs from the MHSA block, these outputs are passed to skip connection as an input to obtain the final output. Before obtaining the final output, the output is fed into the normalization layer and passed to the MLP block. Originally, the MLP block was comprised of linear layers and a GELU activation function. Due to the significant progress in ViT, a locality mechanism [3] is introduced in MLP to capture local features. Furthermore, depth-wise convolution is embedded in the MLP block after the first FC layer to reduce parameters and achieve superior results. Finally, the outcome of the MLP block feeds into skip connection to obtain final outcome from an encoder layer.

4

3. ViTs for Image Classification

Recent development in ViTs have achieved valuable results in the field of classification tasks, several researchers have investigated further models for image classification tasks in different fields such as in medical images, generative images, etc (See in Table. **??**). Where transformers learn a useful representation from natural images. After a huge success of the top-performing transformers in NLP (e.g BERT, GPT, etc). [2] applied transformer for CV tasks and obtained SOTA results compare to conventional CNN and required less computation than CNN. To adjust the image as a sentence, the image is reshaped into 2D flattened patches. Then, learnable embeddings were added to embedded patches same as BERT class tokens. Finally, trainable positional embeddings are added into patch representation to preserve positional information. Transformers are totally dependent on the self-attention mechanism without the use of a convolutional layer. This self-attention technique helps the transformer to understand the relationship between image patches. In theory, an MLP performs better than a CNN model. But data has been a large obstacle with regard to the performance of MLP models. The timeline of ViT methods for image classification task is mentioned in Fig 3)

ViTs [2] models require the training of millions of images and large computational costs, which restricts their use. [4] proposed a competitive visual transformer that only trained on 1.2 million images Image-Net dataset regardless of external data. The original architecture of ViT [2] was modified to enable the native distillation process and adapted the teacher-student learning approach, especially to transformers. In this process, the output of a first neural network (student) is learned from the second neural network (teacher) output.

In addition to this, a distillation token has been added to the transformer, which interacts with image component tokens and classification vectors. In CNN architectures, we can easily increase performance by stacking more convolutional layers but transformers are different which are quickly saturated when architecture becomes deeper. The reason is that as the transformer enters the
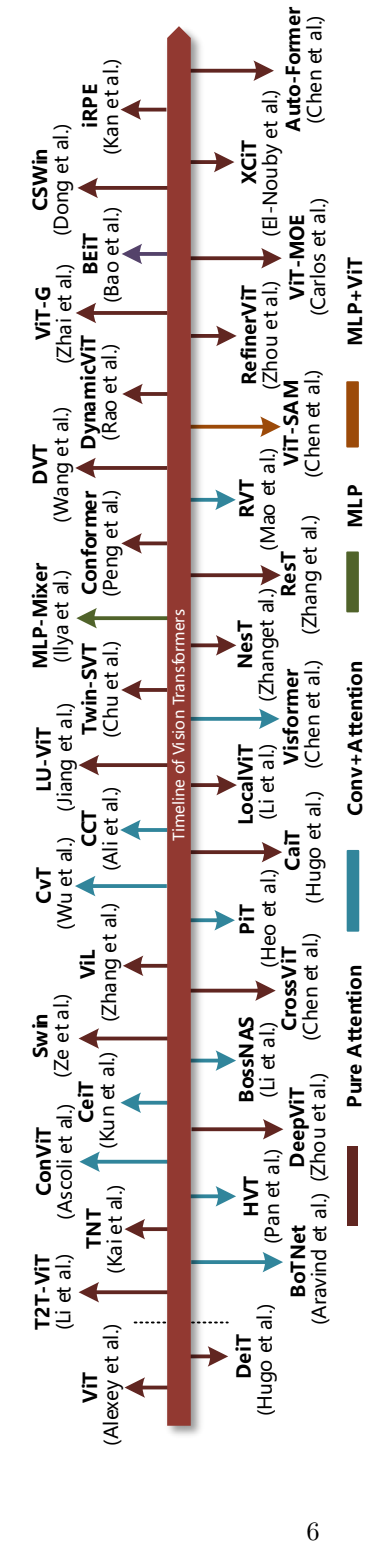
Figure 3: Timeline of existing vision transformer studies including (pure-attention, convolutional with attention, multi-layer perception, vision transformer with multi-layer perception, and self-supervised ViTs). As presented in the figure, the progress has started from the last quarter of 2020. ViT [2] and DeiT [4] is available to the community for 2020. In 2021, a lot of vision transformers have emerged enhancing training strategies, attention mechanism,s and positional encoding techniques. Major contributions to the Image-Net dataset for classification task is shown in this figure.

deep layer, the attention map becomes more and more similar. Based on this, [5] introduced a re-attention module, which regenerated the attention map with a little computational cost in order to enhance the diversity between layers. Then, re-attention module trained on a 32-layer of ViT [2] transformer, the model achieved top-1 accuracy on Image-Net with the increment 1.6% performance. [4] focused on the optimization part in transformers and optimized deeper transformers namely CaiT. It is similar to encoder and decoder where [6] explicitly split transformer layers involved self-attention between patches, from class-attention layers which were dedicated to extracting the content of the processed patches into a single vector so that it could be fed to a linear classifier.

Prior to existing transformers [4, 2, 7], [8] focused on training ViT from scratch and re-designed the tokens-to-token process which is helpful to simultaneously model images based on local-structure information and global correlation. Moreover, the above method significantly reduced the depth and hidden layer dimension. [9] introduced dual-branch ViT to extract multi-scale feature representations and developed a cross-attention-based token-fusion mechanism, which is linear in terms of memory and computation to combine features at different scales.

4. Transformers for Segmentation

Segmentation methods have been greatly increased due to the large-margin success of deep learning methods. It always remains an essential and important topic in the computer vision community. It includes three fundamental segmentation types such as (semantic, instance, and panoptic segmentation). Recently, various transformer methods have shown outstanding results in the domain of segmentation.

4.1. Semantic Segmentation

[20] introduced a sequence-to-sequence approach and replaced the conv-encoder with a pure transformer. Three different decoders are designed to per-

Table 1: Major contribution of vision transformers in different domains of computer vision applications.

Ref	Applications	Highlights
[10]	Object Detection	Fast training convergence and multi-head cross-attention
[11]	Generative Video	Tackling quadratic cost and next-frame prediction
[12]	Object re-identification	Pure vision transformer and person and vehicle re-ID
[13]	Image Restoration	Local-enhanced window and skip-connection schemes
[14]	Action Recognition	Inter-frame attention and mutual-attention fusion
[15]	3D Object Detection	Point clouds backbone and local and global context
[16]	Medical Image Segmentation	Convolution for features and long-range association
[17]	Object Goal Navigation	Object instances in scene and spatial locations, regions
[18]	Tracking	Template, search region and ego-context block
[19]	Visual Recognition	Light-weight attention and two-stage architecture
[19]	Lane Shape Prediction	Non-local interactions and directly regressed output
[20]	Semantic Segmentation	Sequence-to-sequence and different decoder designs
[21]	Generative Adversarial	Bipartite structure and high-resolution synthesis
[10]	Pixel-Wise Prediction	Gates attention approach and multi-scale information
[22]	Deepfake Detection	Mixed architecture and local and global details
[23]	HR Image Generation	Global positioning and double attention
[24]	3D Point Cloud	Light-weight local-attention and 136 times faster inference
[25]	6D object Pose Estimation	Pixel-wise appearance and point-wise geometric
[26]	3D dance Generation	Cascading kinematics and phantom Dance dataset
[27]	Makeup Transfer and Removal	Semantic correspondence and semantic alignment
[28]	Scene Flow Estimation	Irregular point cloud and rich contextual information
[29]	Place Recognition	Super light-weight network and long-range contextual data
[30]	Image Compression	Lightweight image encoder and long-term information
[31]	Fine-Grained Recognition	Part selection module and integrate raw attention weights
[32]	Audio and Speech	Audio spectrogram and generative pre-training
[33]	Crowd Counting	Multi-scale receptive fields and extract semantic features
[34]	Video Restoration	Joint motion estimation parallel feature warping
[35]	Video Grounding	Explore-and-match model and multiple sentences at once
[36]	Video Recognition	Multi-resolution context and lateral connections

form pixel-wise segmentation using progressive sampling, multi-level feature aggregation, and a naive up-sampling approach. [37] presented a convolution-free end-to-end trainable encoder and decoder that captures contextual information. The encoding portion is built on standard ViT and depends on the encoding of patches. Further, a point-wise linear decoder is applied and decoded with a mask transformer. [38] presented SegFormer, a simple yet powerful method with lightweight MLP decoders. An encoder is based on a hierarchical structure that gives multi-scale features and does not require any positional encoding scheme. SegFormer gets rid of complex decoders combining local attention and global attention to gender representation. [39] proposed a fully transformer network which relies on pyramid group transformer encoder to progress learned hierarchical features, while reducing the computational cost of standard ViT. Also, feature pyramid transformer fused spatial-level information and semantic level of the pyramid group transformer.

4.2. Medical Image Segmentation

Medical image segmentation is an important task for the development of healthcare frameworks, particularly for disease diagnosis (e.g biomedical, multi-organ, skin lesion, pannuke and etc). The majority of medical image segmentation problems, the U-NET[1] architectures have become the major standard and achieved top-performing results. Different methods incorporate CNN approach to the self-attention approach to learning long-range dependencies among patches and the convolutional layer is used for extracting feature maps. [40] proposed a hybrid TransUNet for learning high-resolution spatial features from CNN and long-range dependencies from Transformers. TransUNet obtained top-performing performance compared to several FCN-based methods included CNN and self-attention approaches. Transformers are data-hungry and require massive pre-training datasets to achieve comparable results to CNN. Especially for training on small-scale datasets, [41] introduced a gated position-sensitive

---

[1]**U-NET** architecture is a de-facto standard for Medical Image Segmentation

axial attention approach that worked on smaller datasets. Further, they proposed a local-global training technique for transformers. The global part worked on the actual resolution of an image, and a local part operated on image patches.

For 3D medical image segmentation, [42] introduced a volumetric transformer and proposed a unique architecture UNETR. They redefined 3D segmentation as 1D sequence-to-sequence task and utilized a attention mechanism in an encoder to obtain contextual knowledge from patches. Extracted features from encoder is combined with a decoder using residual learning (e.g., like ResNet) at multiple resolutions to predict the segmentation outputs. Similarly to multi-scale architecture, [43] proposed a multi-scale attention technique to capture long-range relation by worked on multi-resolution images. To extract features, CNN-based pyramid architecture is employed. An adaptive-partitioning technique is applied to maintain informative relations and accessed different receptive fields. Recently, various transformer studies have been utilized self-attention mechanism similar to ViT [2]. As a backbone, it is appealing study to employ hierarchical structure in ViT. Swin Transformer [44] is a backbone network for vision tasks based on shifted windows technique. [45] proposed a Swin-Unet, an encoder utilized self-attention technique from local to global and decoder up-sampled the input global features to correspond pixel-level segmentation prediction. Different from previous work, [46] presented Segtran for endless receptive fields to contextualize features for transformer. Segtran is based on Squeeze-and-Expansion an approach that design for solving images tasks. Segtran can easily see the fine details and local images.

5. Transformers for 3D Point Clouds

3D Point cloud aims to collect data in a space with rich geometric information, shape, and scale knowledge. Each data in point cloud data is represented as a 3D object or shape and contains Cartesian coordinates (X, Y, Z). The point collect data is collected from 3D sensors and devices or photogrammetry software. Recently, self-attention-based point clouds methods have been
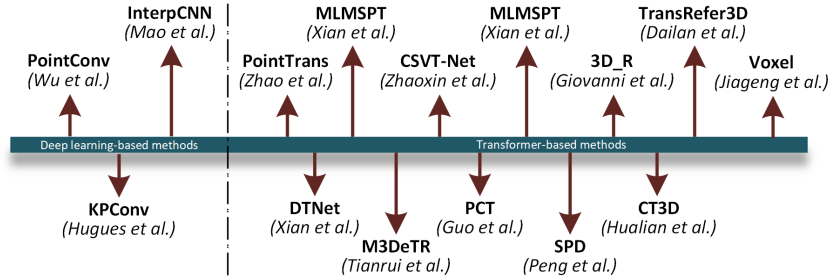
Figure 4: Point Transformer can serve as the backbone for various 3D point cloud tasks. [47]

gained significant importance due to long-range dependencies, especially in this year. Different point cloud approaches have been proposed and achieved SOTA results compare to deep learning methods. Recently, [47] designed an impressive layer to serve as a backbone of dense prediction tasks and scene understanding. A proposed layer is invariant to permutation and cardinality and is therefore inherently suitable for the processing of point cloud data. By utilizing the proposed layer, they constructed a point transformer that is reliable to process point cloud data. Capturing long-range relationships is difficult in CNN networks, transformer solves the long dependencies problem through a self-attention mechanism. The transformer based SOTA method is highlighted in 5. Further, [15] presented a feature learning-based backbone to capture global and local context. The main focus on the local region of the same instance then extends the attention process to other regions. Finally, attended to points from other instances globally, influencing both local and global relationships. [48] developed a DPCT transformer network that can easily capture the long-range relationship and contextual knowledge using point-wise and channel-wise relationships. A unique multi-scale and multi-level point pyramid transformer [49] are designed to capture the cross-level and cross-scale context-aware features.

6. Transformer for Person Re-Identification

The goal of Re-ID is to retrieve a probe person from gallery images over various non-overlapping surveillance cameras. With the progressive development
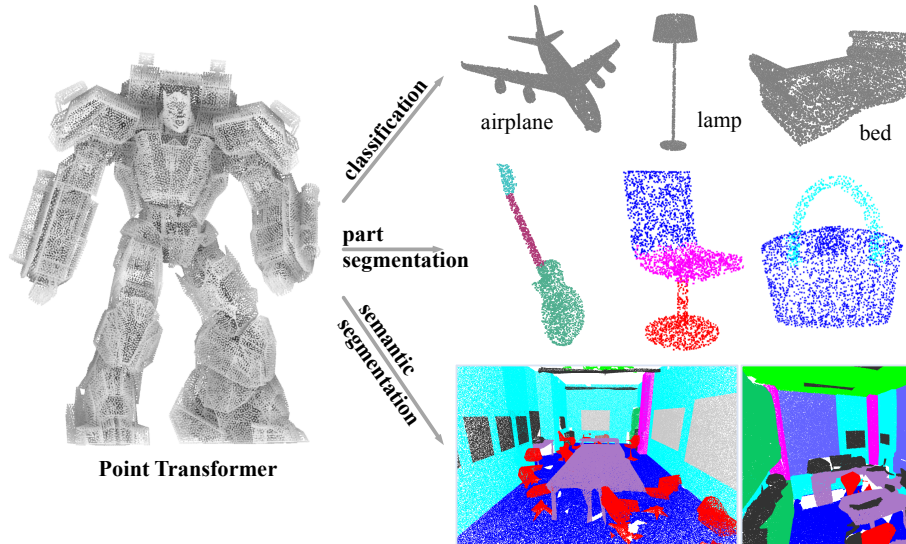
11

Figure 5: Top-performing method for point cloud registration (e.g., classification, part segmentation, and semantic segmentation.)

of DDNs and the rising requirement of high-tech video surveillance cameras, it has gained great interest in the CV and PR community. Re-ID has been broadly considered a topic in research as a particular person retrieval problem over non-overlapping CCTV cameras. Providing a probe image, the target of the Re-ID task is to identify whether this human has seemed in another place at a different time span captured by a different CCTV camera, or even the same camera at a different time instant. The probe person can be represented by an image, a video sequence, and even a text description.

Extracting discriminative features is a major and key problem in Re-ID. While CNNs based approaches have obtained remarkable success, they only process information via fixed-size grid mechanism at one time and lose important information details due to convolution and down-sampling operators (e.g. pooling and stride). Several transformer-based Re-ID methods have been designed to overcome these problems, [12] presented a strong baseline that utilized pure transformer and designed a jigsaw patches module, containing shift and patch shuffle operation, which promotes perturbation-invariant and robust dis-
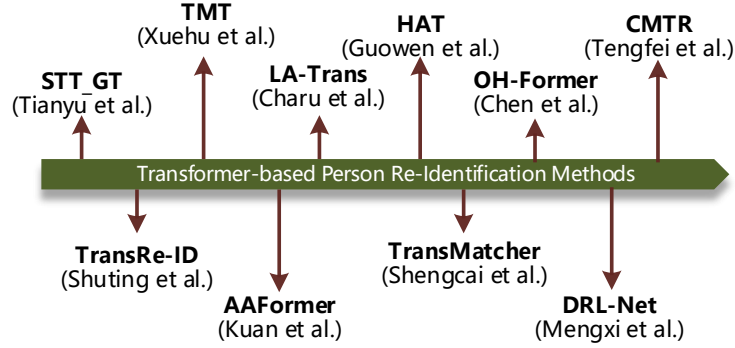
12

Figure 6: Timeline of self-attention-based mechanism methods for person re-identification.

criminative features. Further, they introduced an SIE module that encoded side information via learnable embeddings. Xuehu et a. [50] proposed a pure transformer-based Re-ID method to obtain rich features and extract in-depth video representation. Particularly, to jointly transform raw videos into spatial and temporal domains, a trigeminal feature extractor is proposed. Inspired by transformer architecture, three self-view transformers are designed to utilize token's dependencies and local information in ST domains. For extensive video representations, a cross-view method is developed to combine the multi-view features.

Different from Re-ID, video re-ID consists of continuous frames from surveillance cameras providing more rich spatio-temporal information that can be more effective and carefully note the movement of identity, view angle-changes, and appearance. [51] presented a simple transformer-based re-ID framework for the video re-ID task. Particularly, the constrained attention technique and global attention learning module are designed to utilize ST information from videos. To minimize the possibility of over-fitting, they introduced a synthesized data pre-training strategy to a better-initialized framework. In re-ID, extracting part-level features is a critical task to design a robust framework. [52] introduced an alignment technique in transformer and presented a transformer to automatically find person and non-person parts at a single time at the patch level. Recently, in part-level works, [53] combined the output of global and local

13

token of ViTs with a PCB-like technique to improve re-ID performance. They further incorporated a block-wise fine-tuning scheme to regularize the fine-tuned of a pre-trained backbone of ViTs. [54] applied a transformer for image matching and metric learning and proposed a simplified decoder for efficient image matching with an emphasis on similarity computation and mapping.

Combining CNNs and transformers is the new direction in the computer vision community and achieves more impressive results. Guowen e al. [55] utilized CNNs and transformers to design a re-ID framework and proposed the DSA method to enhance the multi-scale features and combine hierarchical features from the CNN backbone. To integrate low-level information, transformer feature calibration is employed to obtain prominent knowledge and feed into each level of hierarchical features. Similarly, part-aware transformer [56] adopted CNN as a backbone network and proposed an occluded re-ID method. Encoder-decoder architecture is presented to utilize different portions of the human body. Part diversity and part discriminability are designed to obtain part-prototypes well with identity labels. [57] proposed disentangled representation network which handled occluded re-ID using global reasoning of local features without needing alignment of person image or further guidance. It's automatically measured image similarity of a person's body parts under the supervision of semantic knowledge of objects.

6.1. Future Prospects

Although recent years have observed remarkable progress in ViTs, there is still a lot of research that is hidden and unexplored. In this section, we highlight several new research directions for future work.

Most of the proposed ViT variants are highly computational expensive, require significant hardware, and are time-consuming. The first direction is to apply CNN pruning principle to prune ViTs according to their learnable scores. This technique will be helpful to reduce the complexity and memory footprint of ViT. In terms of the inference speed of ViT, pruning useless features, and preserving important features so that the final model can run more efficiently.

Thus, deployment of the model becomes easier in real-time devices (e.g., smartphone, surveillance system) while maintaining the low resource cost. Therefore, it is a new research direction to obtain top-performing results while reducing the number of parameters and FLOPS.

There are several unique qualities are present in pure ViT models like self-attention, global context, and generalization. While CNN is a widely used technique and popular for extracting pixel-level features (local features) and shared weights. Interestingly, the majority of studies have been investigated that a combination of CNN and ViT is more worthy instead of using straight-forward pure ViTs. The main reason behind this, pure ViTs have lack shared weights, local context, and receptive fields. On other hand, CNN only works on pixel-level features, compressed all information through filters and does not consider the distanced pixel-relationship. Thus, it is an interesting topic to develop convolutional ViT or develop pure ViT without CNN, this question is still remains unsolved and more study required in this domain.

ViT architectures have achieved superior performance compared to CNN in difficult tasks such as dense prediction and tiny object detection. These ViT models can easily able to learn an internal representation of the visual data which can surpass human capability. To some extent, an internal representation of the visual data remains secret in the architecture and keeps this technique as a black box. To explain how it works from an internal side, it is better to develop a new visualization layer that aims to understand how it works and be able to interpret its results. Thus, it is an interesting research area to develop an explainable vision transformer.

Positional encoding is an important part of a transformer to capture a sequence of input tokens. However, positional encoding is not well investigated and studied, always remains a controversial topic which encoding scheme is best for it (e.g., conditional encoding or absolute encoding). Currently, the evidence available in the [58] relative positional encoding scheme is able to achieve top-performing results and focus on positional directivity. This question is for the future whether we need more positional encoding schemes or not.

7. Conclusion

ViTs are gaining more popularity and showing more impressive results in the domain of CV due to the advantage of the self-attention mechanism and creating a long-range relationship. In this chapter, we first describe the fundamental concepts of vision transformers and review the proposed ViT methods in recent years. Secondly, we highlight the influential work and draw a timeline of ViTs in different domains (e.g., classification, segmentation, point cloud, and person re-identification). Lastly, we comparatively study different ViTs and CNN methods in terms of their accuracy on an image-Net dataset. Overall, ViT methods demonstrate excellent performance, and combining the self-attention mechanism with CNN shows outstanding results. Still, the training strategies and much-hidden work of ViTs have not been discovered yet and fully investigated.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: NeurIPS, 2017.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, ICLR(2021) .

[3] Y. Li, K. Zhang, J. Cao, R. Timofte, L. Van Gool, Localvit: Bringing locality to vision transformers, arXiv(2021) .

[4] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, ICML(2020) .

[5] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Q. Hou, J. Feng, Deepvit: Towards deeper vision transformer, arXiv(2021) .

[6] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, H. Jégou, Going deeper with image transformers, ICCV(2021) .

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: **ECCV**, 2020.

[8] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, ICCV(2021) .

[9] C.-F. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, ICCV(2021) .

[10] G. Yang, H. Tang, M. Ding, N. Sebe, E. Ricci, Transformer-based attention networks for continuous pixel-wise prediction, ICCV(2021) .

[11] Y.-F. Wu, J. Yoon, S. Ahn, Generative video transformer: Can objects be the words?, in: ICML, 2021.

[12] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, Transreid: Transformer-based object re-identification, arXiv(2021) .

[13] Z. Wang, X. Cun, J. Bao, J. Liu, Uformer: A general u-shaped transformer for image restoration, arXiv(2021) .

[14] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, W. Li, Trear: Transformer-based rgb-d egocentric action recognition, IEEE Transactions on Cognitive and Developmental Systems(2021) .

[15] X. Pan, Z. Xia, S. Song, L. E. Li, G. Huang, 3d object detection with pointformer, in: **CVPR**, 2021.

[16] Y. Gao, M. Zhou, D. Metaxas, Utnet: A hybrid transformer architecture for medical image segmentation, MICCAI(2021) .

[17] H. Du, X. Yu, L. Zheng, Vtnet: Visual transformer network for object goal navigation, in: In ICLR, 2020.

[18] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, H. Lu, Transformer tracking, in: **CVPR**, 2021.

[19] L. Yuan, Q. Hou, Z. Jiang, J. Feng, S. Yan, Volo: Vision outlooker for visual recognition, arXiv(2021) .

[20] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: **CVPR**, 2021.

[21] D. A. Hudson, C. L. Zitnick, Generative adversarial transformers, ICML(2021) .

[22] D. Coccomini, N. Messina, C. Gennaro, F. Falchi, Combining efficientnet and vision transformers for video deepfake detection, arXiv(2022) .

[23] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, B. Guo, Styleswin: Transformer-based gan for high-resolution image generation, arXiv(2021) .

[24] C. Park, Y. Jeong, M. Cho, J. Park, Fast point transformer, arXiv(2021) .

[25] L. Zou, Z. Huang, N. Gu, G. Wang, 6d-vit: Category-level 6d object pose estimation via transformer-based instance representation learning, arXiv(2021) .

[26] B. Li, Y. Zhao, Z. Shi, L. Sheng, Danceformer: Music conditioned 3d dance generation with parametric motion transformer, AAAI(2022) .

[27] Z. Sun, Y. Chen, S. Xiong, Ssat: A symmetric semantic-aware transformer network for makeup transfer and removal, AAAI(2022) .

[28] B. Li, C. Zheng, S. Giancola, B. Ghanem, Sctn: Sparse convolution-transformer network for scene flow estimation, AAAI(2022) .

[29] Z. Fan, Z. Song, H. Liu, Z. Lu, J. He, X. Du, Svt-net: Super light-weight sparse voxel transformer for large scale place recognition, AAAI(2022) .

[30] Y. Bai, X. Yang, X. Liu, J. Jiang, Y. Wang, X. Ji, W. Gao, Towards end-to-end image compression and analysis with transformers, AAAI(2022) .

[31] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, C. Wang, A. Yuille, Transfg: A transformer architecture for fine-grained recognition, AAAI(2022) .

[32] Y. Gong, C.-I. J. Lai, Y.-A. Chung, J. Glass, Ssast: Self-supervised audio spectrogram transformer, AAAI(2022) .

[33] Y. Tian, X. Chu, H. Wang, Cctrans: Simplifying and improving crowd counting with transformer, AAAI(2022) .

[34] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, L. Van Gool, Vrt: A video restoration transformer, AAAI(2022) .

[35] S. Woo, J. Park, I. Koo, S. Lee, M. Jeong, C. Kim, Explore and match: End-to-end video grounding with transformer, AAAI(2022) .

[36] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun, C. Schmid, Multiview transformers for video recognition, arXiv(2022) .

[37] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: Transformer for semantic segmentation, ICCV(2021) .

[38] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, Segformer: Simple and efficient design for semantic segmentation with transformers, arXiv(2021) .

[39] S. Wu, T. Wu, F. Lin, S. Tian, G. Guo, Fully transformer networks for semantic image segmentation, arXiv(2021) .

[40] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, arXiv(2021) .

[41] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, V. M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, arXiv(2021) .

[42] A. Hatamizadeh, D. Yang, H. Roth, D. Xu, Unetr: Transformers for 3d medical image segmentation, WACV(2022) .

[43] Z. Zhang, B. Sun, W. Zhang, Pyramid medical transformer for medical image segmentation, arXiv(2021) .

[44] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, ICCV(2021) .

[45] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, arXiv(2021) .

[46] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, R. S. M. Goh, Medical image segmentation using squeeze-and-expansion transformers, IJCAI(2021) .

[47] H. Zhao, L. Jiang, J. Jia, P. Torr, V. Koltun, Point transformer, **ICCV**(2021) .

[48] X.-F. Han, Y.-F. Jin, H.-X. Cheng, G.-Q. Xiao, Dual transformer for point cloud analysis, arXiv(2021) .

[49] X.-F. Han, Y.-J. Kuang, G.-Q. Xiao, Point cloud learning with transformer, arXiv(2021) .

[50] X. Liu, P. Zhang, C. Yu, H. Lu, X. Qian, X. Yang, A video is worth three views: Trigeminal transformers for video-based person re-identification, arXiv(2021) .

[51] T. Zhang, L. Wei, L. Xie, Z. Zhuang, Y. Zhang, B. Li, Q. Tian, Spatiotemporal transformer for video-based person re-identification, arXiv(2021) .

[52] K. Zhu, H. Guo, S. Zhang, Y. Wang, G. Huang, H. Qiao, J. Liu, J. Wang, M. Tang, Aaformer: Auto-aligned transformer for person re-identification, arXiv(2021) .

[53] C. Sharma, S. R. Kapil, D. Chapman, Person re-identification with a locally aware transformer, arXiv(2021) .

[54] S. Liao, L. Shao, Transformer-based deep image matching for generalizable person re-identification, arXiv(2021) .

[55] G. Zhang, P. Zhang, J. Qi, H. Lu, Hat: Hierarchical aggregation transformers for person re-identification, arXiv(2021) .

[56] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, F. Wu, Diverse part discovery: Occluded person re-identification with part-aware transformer, in: **CVPR**, 2021.

[57] M. Jia, X. Cheng, S. Lu, J. Zhang, Learning disentangled representation implicitly via transformer for occluded person re-identification, arXiv(2021) .

[58] K. Wu, H. Peng, M. Chen, J. Fu, H. Chao, Rethinking and improving relative position encoding for vision transformer, in: ICCV, 2021.