2024 IEEE 10th International Conference on Intelligent Data and Security (IDS) | 979-8-3503-8948-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/IDS62739.2024.00011

A Survey of Applications of Vision Transformer and Its Variants

1st Chuang Wu
School of Computer Science and Engineering,
Hunan University of Science and Technology
Hunan Key Laboratory for Service computing
and Novel Software Technology City, Country
2534765383@qq.com

2nd Tingqin He
School of Computer Science and Engineering,
Hunan University of Science and Technology
Hunan Key Laboratory for Service computing
and Novel Software Technology City, Country
hetingqin@hnust.edu.cn

***Abstract*—The Transformer architecture, renowned for its efficacy in natural language processing, encounters unique hurdles when applied to computer vision. In response, the Vision Transformer (ViT) emerges as a successful adaptation for image classification tasks. While ViT exhibits tremendous potential in revolutionizing computer vision, addressing its inherent challenges and limitations stands as a critical endeavor. This comprehensive survey meticulously scrutinizes the drawbacks associated with ViT, proposing bespoke adaptations tailored to specific applications while showcasing their remarkable performance across diverse visual tasks. Moreover, it delves into the evolution of ViT adaptations across various visual domains, elucidating four promising directions for future research and development in this dynamic field.**

***Index Terms*—Transformer, Computer Vision, Self-attention, High-level vision, Low-level vision**

I. Introduction

A. Introduction to Transformer

Within the domain of natural language processing (NLP) [1]–[3], recurrent neural networks (RNNs) [4] and long short-term memory networks (LSTMs) [5] have historically stood as pivotal model frameworks. Nonetheless, these models grapple with the inherent challenge of intricate parallelization, resulting in elevated computational complexity [15] , [16] , [17]and protracted training durations [24] , [43]. Thankfully, the introduction of the Transformer model has effectively mitigated this issue. Capitalizing on the potency of attention mechanisms, the Transformer model facilitates parallel task processing, catalyzing remarkable advancements in NLP. The ascendancy of influential models such as BERT [6] and GPT [7] serves as compelling testimony to this paradigmatic shift. Subsequently, the ensuing subsection will furnish an exhaustive overview of the original Transformer's architectural framework, intricately dissecting its fundamental components.

Encoder-decoder architecture. The Transformer can be grouped into two parts, as shown in Fig. 1. Encoder is composed of multi-head attention layers and feed-forward neural networks. The decoder has a similar structure to the encoder but with an additional masked multi-head attention mechanism. In addition, in order to make better use of the position information between input elements, Transformer uses position encoding to encode the position information of input elements through trigonometric functions. As shown in (1) and (2) :

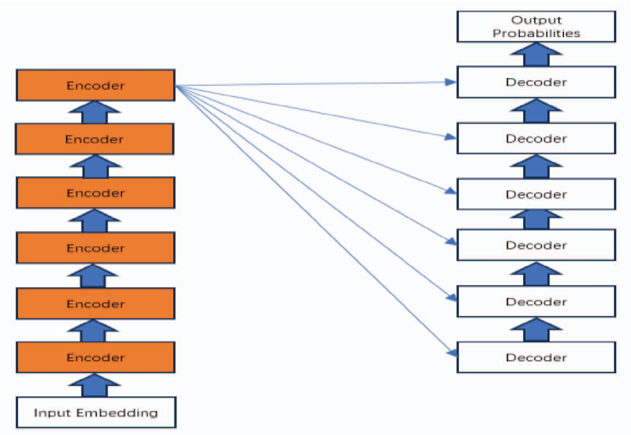

Fig. 1. Transformer model structure.

$$PE_{pos,\ 2i} = \sin\left(pos/\left(10000^{2i/d_{model}}\right)\right) \tag{1}$$

$$PE_{pos,\ 2i+1} = \cos\left(pos\left(10000^{2i/d_{model}}\right)\right) \tag{2}$$

Where pos represents the position in the sequence, i is the dimension of the positional encoding, and d_model is the dimension of the model.

Attention Mechanism. The Transformer structure features a multi-head attention mechanism, which adeptly captures dependencies among positions in the input sequence. By allowing parallelized attention to multiple positions, the model effectively captures correlations between them. It can be formulated as (3) and (4):

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{3}$$

$$MultiHead\left(Q, K, V\right) = Concat\left(head_i, ..., head_h\right)W^o \tag{4}$$

Where Q, K, V respectively represent queries, keys, and values. $concat$ represents the result of splicing all heads. W^o is to keep our input dimension and output dimension

979-8-3503-8948-7/24/$31.00 ©2024 IEEE
DOI 10.1109/IDS62739.2024.00011

consistent.W_i^Q, W_i^K: and W_i^V; are the parameter matrices for linear transformation. Attention is the attention algorithm used in the multi-head attention mechanism [8] .In order to reduce the influence brought by the increase of vector dimension, $\sqrt{d_k}$ is introduced in the dot product attention. The attention formula is as follows (5)

$$Attention\left(Q,K,V\right) = soft\max\left(QK^T/\sqrt{d_k}\right)V \qquad (5)$$

B. Introduction to ViT

The ViT model innovatively applies the Transformer architecture to image classification tasks. It accomplishes this by partitioning the image into fixed-sized patches and deriving vector representations for each patch through linear transformations, mirroring the tokenization process in natural language processing. These patch embeddings, enriched with location information, undergo processing within the Transformer encoder for both feature extraction and classification. Refer to Fig. 2 for a visual depiction of this sequential process.

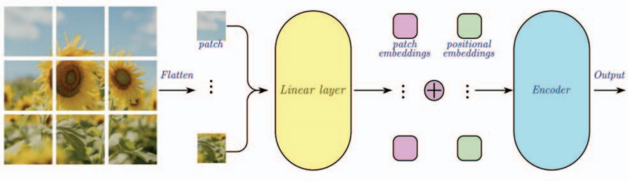

Fig. 2. The process of processing pictures in ViT

The architecture of the ViT model primarily comprises two integral components: image patch-based feature extraction and Transformer-based feature encoding. The former involves the precise segmentation of the image into uniformly sized patches, which are subsequently transformed into patch embeddings. The latter intricately processes these embeddings to generate a comprehensive global feature representation. Ultimately, these refined global features seamlessly integrate into a fully connected layer, culminating in precise classification and yielding the desired prediction outcomes.

II. The Development of Vision transformer

The ViT model excels in image classification on smaller datasets, particularly when trained on large datasets. It can process images of varying sizes and integrates with models like segmentation networks and generative models [14]. This section discusses ViT's limitations and the enhancements made to address them, also showcasing these updated versions' performance in various settings.

A. Disadvantages of ViT and its related variants

In line with the ViT paradigm, a series of ViT variants have been proposed to enhance the performance of visual tasks [9]–[13]. These variants primarily tackle various challenges, such as the high computational cost associated with ViT, the difficulty in capturing low-level information, the limitations in stacking deeper layers, and the constraints of utilizing ViT in resource-limited devices. Furthermore, we provide a concise summary of the performance of each variant on the ImageNet-1k, COCO2017, and ADE20K datasets, showcasing their effectiveness. As shown in Tab. I.

TABLE I
The performance of various variants of ViT on different visual tasks.

	Top-1(%)	mloU(%)	AP
PVT	81.7	44.8	40.4
Swin-Trans-former	84.2	46.1	
Dual-ViT	85.7	49.2	47.4
Slide-Trans-former	84.9	48.5	46.8
DeiT	84.5	-	-
STViT	86.4	47.6	48.6
OnA-ViT	84.8	-	38.5
Shunted-ViT	84.0	48.2	47.1
DeepViT	80.9	-	-
CaiT	86.5	-	-
WideNet	80.1	-	-
MobileViT	78.4	-	-
EifficientViT	77.1	32.7	-
Topformer	-	-	33.4
RepViT	81.4	42.8	-

High computational complexity. The ViT model employs self-attention for image processing but struggles with high computational demands. Wang et al. [18] addressed this in the PVT model by integrating a pyramid structure and Patch Embedding to reduce input resolution. This model also effectively shortens the length of K and V, balancing resolution and field. Liu et al. [19] introduced the Swin Transformer, utilizing a shifted window mechanism to enhance global modeling. Mei Tao et al. [20] developed the Dual ViT model, which efficiently merges global and local semantic features. Traditional methods like Linformer [21] and Longformer [22] face limitations in learning local features. Gao et al. [23] proposed the Slide-Transformer, reinterpreting the lm2Col function and incorporating a deformation shift module for efficient and flexible local self-attention.

Difficult to obtain underlying information. The ViT model, while adept at longrange dependencies, struggles with local feature extraction. The DeiT model [8] addresses this by introducing a teacher-student distillation strategy and token-based distillation, enhancing ViT's efficiency and accuracy. Huang et al. [25] further developed this with the STViT model, which utilizes sparse associations to predict super tokens and self-attention in the super token space, improving long-term dependency mapping. Moab Arar et al. [11] introduced the QnA-ViT model, offering a local attention layer with linear complexity and shift invariance, optimizing memory usage with shift-invariant local attention layers. Lastly, Ren et al. [10] proposed Shunted-ViT, integrating a Shunt Split Attention (SSA) mechanism in self-attention heads to diversify feature consideration and scale modeling.

The limitations in stacking layers. As ViT model layers increase, task performance may suffer. Zhou et al. [26] found that deep ViT's reduced efficacy is due to similar attention maps in deep blocks, addressed by adding a feature transformation matrix and normalization to the self-attention mechanism. Hugo Touvron et al. [27] created the CaiT model, improving performance with LayerScale, a diagonal matrix enhancing each residual block's output, and Class-Attention, focusing on token information extraction. Additionally, Xue et al. [28] enhanced the Transformer model by widening it, using a Mixture of Experts (MoE) layer shared across blocks, allowing unique token representations and independent normalization for each block, making WideNet more efficient with fewer parameters.

The challenges of applying on resource-constrained devices. ViT, though powerful, faces challenges on resource-constrained devices. Addressing this, researchers developed lightweight ViT models for mobile applications. Mehta et al. [29] introduced MobileViT, combining CNN with Transformer for more stable training and faster convergence. It offers superior performance with fewer parameters than standard ViT. Wang et al. [30] proposed RepViT, enhancing standard lightweight CNNs, and outperforming many lightweight ViT models. Zhang et al. [31] presented Top-Former, a Token Pyramid Vision Transformer for mobile devices, using varied scale tokens to generate enhanced semantic features. Furthermore, Zhang et al. [32] introduced EfficientViT, with its unique block structure and cascaded group attention modules, optimizing memory access and computational efficiency.

III. Applications of Vision Transformer on low-level visual tasks

Low-level vision tasks represent a significant domain within computer vision, focusing on extracting fundamental visual features and structures from pixel-level information in images and videos. The following subsection primarily elucidates video reconstruction and image generation.

A. Super-resolution video reconstruction

Super-resolution video reconstruction (SRVR)enhances low-resolution video quality but struggles with temporal dependencies. TTVSR addresses this by partitioning frames into trajectories of visual tokens, improving resolution via self-attention mechanisms. Similarly, *stereo video super-resolution* (SVSR) aims to increase spatial resolution. facing challenges in maintaining stereo and temporal integrity, potentially causing 3D fatigue. Tran-SVSR [35] overcomes these with spatio-temporal self-attention and optical flow layers, adding a Parallax *Attention Mechanism* (PAM) for better stereo fusion. It outperforms SOTA models on the KITTI2012 [36] and KITTI2015 [37] benchmarks.

B. Image generation

Image generation, a remarkable computer capability, involves various Transformer-based algorithms like Image Transformer and TransGAN [38]. Image Transformer treats this task as a self-autoregressive sequence problem, creating realistic, diverse images, though it demands significant computational power and extensive data, possibly sacrificing detail and texture. TransGAN maps images to vectors using a Transformer encoder and a decoder, evaluated by a pyramidal discriminator. It generates highly realistic images but has stability issues during training. Presently, ViT-based image generation is led by ViTGAN [39] and GAN Inverse Mapping for pre-training Style-GAN [40]. ViTGAN uses ViT as a discriminator, enhanced with gradient penalty and spectral normalization for stable training, achieving results similar to CNN-based StyleGAN2. The other method involves mapping noise vectors to W+ space codes using a multilayer Transformer, allowing for high-quality, low-distortion image generation through flexible code editing.

IV. Application of Vision Transformer on Advanced Vision Tasks

In recent times, there has been an increasing interest in utilizing ViT for advanced computer vision tasks, including object detection, semantic segmentation, and instance segmentation. In this section, we will review these methods.

A. Object detection

Presently, deep learning-based object detection algorithms can be broadly categorized into two distinct types: two-stage detectors and one-stage detectors.Two-stage detectors, as the name implies, encompass two stages: region extraction and classification. Noteworthy examples of such algorithms include RCNN and FastRCNN [41].Conversely, one-stage detectors directly detect objects within the entire image. Prominent examples of this approach include YOLO [42] and SSD. In recent times, there has been a growing interest in Transformer-based object detection algorithms, primarily centered around enhancing the performance of DETR algorithm [9] and addressing the challenges encountered in 3D object detection.

Improvements on the slow convergence of DETR. The DETR algorithm represents a novel approach to object detection. However, its training process encounters a bottleneck due to its sluggish convergence speed.To overcome this challenge, researchers have proposed two methods: SAM-DETR and DN-DETR [44]. SAM-DETR introduces a semantic matching technique that projects object queries into an embedding space identical to the encoded image features. By explicitly seeking out salient points with the most discriminative features for semantic alignment matching, the convergence speed of DETR is significantly accelerated. DN-DETR, on the other hand, presents a novel denoising training approach. It involves inputting noisy ground truth bounding boxes into the Transformer decoder and training the model to reconstruct the original bounding boxes. This effectively mitigates the challenges associated with bidirectional graph matching, leading to accelerated convergence speed.

Improvements on problems in 3D object detection. In autonomous driving, monocular 3D detection is key. MonoDTR [45] stands out with its minimal design and DETR-based architecture, using a Transformer encoder for feature extraction and 3D detection. DA-BEV [46] adopts a bird's eye view for 3D detection, integrating Depth-Aware Spatial Cross-Attention and Depth-wise Contrastive Learning to enhance depth perception. Song et al. [47] introduced ViDT, using a Swin Transformer with a Reconfigured Attention Module for scalable object detection and an encoder-less neck to reduce computational load.

B. Image segmentation

In computer vision, semantic segmentation categorizes each pixel, whereas instance segmentation separates pixels of the same category into distinct objects.

Semantic segmentation. In computer vision, semantic segmentation requires precise pixel-level labels for training, but obtaining these can be costly. This has led to increased research in weakly supervised semantic segmentation (WSSS). Current methods largely use CNNs, but integrating ViT into WSSS is emerging. However, ViTbased class activation maps (CAM) as affinity labels are often inaccurate and incomplete. Ru et al. [48] introduced the Affinity from Attention (AFA) module, utilizing Transformer's multi-head self-attention to learn semantic affinity, later used as a pseudo label. They also developed the Pixel-Adaptive Refinement module, integrating image color and spatial information. Furthermore, Xu et al. proposed MCTformer, leveraging interactions between class and patch tokens to create class-discriminative object localization maps, complementing CAM and improving pseudo-label quality for WSSS. H. Qiu et al. [49] proposed deep residual learning-based enhanced JPEG compression in the Internet of Things [50]–[52], and integrated AI security [53], [54] within the framework, such as the back door attack and the adversary attack [55]–[57] to video systems.

Instance segmentation. Instance segmentation, a key task in computer vision, involves delineating objects in images and creating masks for each. Current methods fall into two-stage and query-based segmentation. A new method, Mask Transfiner, detects and decomposes image regions into a multi-level quadtree, transforming points into query sequences for a Transformer model to predict final labels. This approach offers better mask quality with lower computational and memory requirements compared to CNN-based methods. Additionally, instance segmentation extends to videos, which contain spatial and temporal data. Wang et al. proposed VisTR, an end-to-end parallel sequence decoding approach for video instance segmentation, processing multiple image frames as a video segment and directly outputting masks for each instance.

V. Conclusion

This study rigorously investigated the ViT and its diverse variants within the domain of computer vision, comprehensively addressing their emergent applications alongside inherent challenges. While ViT's capacity to uniformly process image blocks showcases promise, its limitations become apparent, particularly in its ability to navigate a spectrum of tasks and elucidate model decisions within computer vision. This article advocated for augmenting performance through targeted attention mechanisms focused on pivotal image tokens, thereby amplifying interpretability. Moreover, in confronting the challenges posed by limited datasets, it proposes robust solutions such as employing data augmentation techniques encompassing rotation, cropping, and scaling to fortify model resilience. Furthermore, the discourse focuses on optimizing ViT by strategically weighting influential patches, emphasizing key objects and pertinent information within the visual context. Notably, the article underscores the critical need for real-time processing efficiency, advocating for model acceleration methodologies leveraging lightweight network architectures or dedicated hardware such as GPUs to meet the stringent demands imposed by real-time applications.

Acknowledgment

This research was funded by the Hunan Provincial Natural Science Foundation of China(2021JJ40214) and the Scientific Research Fund of Hunan Provincial Education Department (Nos.20C0781).

References

[1] G. Vaswani, A. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., and others, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[4] N. Peng and M. Dredze, "Improving named entity recognition for Chinese social media with word segmentation representation learning," arXiv preprint arXiv:1603.00786, 2016.

[5] R. Xiong, C. Wu, L. Xiong, and L. Li, "Character-level Based Conference Named Entity Recognition Using Bi-LSTM," in 2019 International Conference on Computer, Network, Communication and Information Systems (CNCI 2019), Atlantis Press, 2019, pp. 83-88.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[7] T. Brown et al., "Language models are few-shot learners," in Advances in neural information processing systems, vol. 33, pp. 1877-1901, 2020.

[8] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in International conference on machine learning, PMLR, 2021, pp. 10347-10357.

[9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European conference on computer vision, Springer, 2020, pp. 213-229.

[10] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10853-10862.

[11] M. Arar, A. Shamir, and A. H. Bermano, "Learned queries for efficient local attention," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10841-10852.

[12] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, "Mobile-former: Bridging mobilenet and transformer," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5270-5279.

[13] T.-J. Fu, X. E. Wang, S. T. Grafton, M. P. Eckstein, and W. Y. Wang, "Language-based video editing via multi-modal multi-level transformer," arXiv preprint arXiv:2104.01122, vol. 1, 2021.

[14] I. Goodfellow et al., "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[15] W. Liang, K.-C. Li, J. Long, X. Kui, and A. Y. Zomaya, "An industrial network intrusion detection algorithm based on multifeature data clustering optimization model," IEEE Transactions on Industrial Informatics, vol. 16, no. 3, pp. 2063-2071, 2019.

[16] W. Liang, M. Tang, J. Long, X. Peng, J. Xu, and K.-C. Li, "A secure fabric blockchain-based data transmission technique for industrial Internet-of-Things," IEEE Transactions on Industrial Informatics, vol. 15, no. 6, pp. 3582-3592, 2019.

[17] W. Liang, D. Zhang, X. Lei, M. Tang, K.-C. Li, and A. Y. Zomaya, ”Circuit copyright blockchain: blockchain-based homomorphic encryption for IP circuit protection,” IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 3, pp. 1410-1420, 2020.

[18] W. Wang et al., ”Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 568-578.

[19] Z. Liu et al., ”Swin transformer: Hierarchical vision transformer using shifted windows,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012-10022.

[20] T. Yao et al., ”Dual vision transformer,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[21] Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768.

[22] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

[23] X. Pan, T. Ye, Z. Xia, S. Song, and G. Huang, ”Slide-transformer: Hierarchical vision transformer with local self-attention,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 2082-2091.

[24] Z. Xu, W. Liang, K.-C. Li, J. Xu, A. Y. Zomaya, and J. Zhang, ”A time-sensitive token-based anonymous authentication and dynamic group key agreement scheme for industry 5.0,” IEEE Transactions on Industrial Informatics, vol. 18, no. 10, pp. 7118-7127, 2021.

[25] H. Huang, X. Zhou, J. Cao, R. He, and T. Tan, ”Vision transformer with super token sampling,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22690-22699, 2023.

[26] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, ”Deepvit: Towards deeper vision transformer,” arXiv preprint arXiv:2103.11886, 2021.

[27] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, ”Going deeper with image transformers,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 32-42.

[28] F. Xue, Z. Shi, F. Wei, Y. Lou, Y. Liu, and Y. You, ”Go wider instead of deeper,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 8, pp. 8779-8787, 2022.

[29] S. Mehta and M. Rastegari, ”Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer,” arXiv preprint arXiv:2110.02178, 2021.

[30] A. Wang, H. Chen, Z. Lin, H. Pu, and G. Ding, ”Repvit: Revisiting mobile CNN from ViT perspective,” arXiv preprint arXiv:2307.09283, 2023.

[31] W. Zhang, Z. Huang, G. Luo, T. Chen, X. Wang, W. Liu, G. Yu, and C. Shen, ”Topformer: Token pyramid transformer for mobile semantic segmentation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12083-12093.

[32] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, ”Efficientvit: Memory efficient vision transformer with cascaded group attention,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14420-14430.

[33] Y. Liu, W. Liang, K. Xie, S. Xie, K. Li, and W. Meng, ”LightPay: A Lightweight and Secure Off-Chain Multi-Path Payment Scheme Based on Adapter Signatures,” IEEE Transactions on Services Computing, 2023.

[34] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, ”Basicvsr: The search for essential components in video super-resolution and beyond,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4947-4956.

[35] H. Imani, M. B. Islam, and L.-K. Wong, ”A new dataset and transformer for stereoscopic video super-resolution,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 706-715.

[36] A. Geiger, P. Lenz, and R. Urtasun, ”Are we ready for autonomous driving? The KITTI vision benchmark suite,” in 2012 IEEE conference on computer vision and pattern recognition, 2012, pp. 3354-3361.

[37] M. Menze and A. Geiger, ”Object scene flow for autonomous vehicles,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3061-3070.

[38] Y. Jiang, S. Chang, and Z. Wang, ”Transgan: Two pure transformers can make one strong gan, and that can scale up,” Advances in Neural Information Processing Systems, vol. 34, pp. 14745-14758, 2021.

[39] Y. Gunduc, ”Vit-GAN: Image-to-image Translation with Vision Transformers and Conditional GANs,” Authorea Preprints, 2023.

[40] X. Hu, Q. Huang, Z. Shi, S. Li, C. Gao, L. Sun, and Q. Li, ”Style transformer for image inversion and editing,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11337-11346.

[41] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and others, ”Sparse R-CNN: End-to-end object detection with learnable proposals,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14454-14463.

[42] X. Han, J. Chang, and K. Wang, ”You only look once: unified, real-time object detection,” Procedia Computer Science, vol. 183, no. 1, pp. 61-72, 2021.

[43] J. Long, W. Liang, K.-C. Li, Y. Wei, and M. D. Marino, ”A regularized cross-layer ladder network for intrusion detection in industrial Internet of Things,” IEEE Transactions on Industrial Informatics, vol. 19, no. 2, pp. 1747-1755, 2022.

[44] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, ”DN-DETR: Accelerate DETR training by introducing query denoising,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 13619-13627.

[45] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, ”MonoDTR: Monocular 3D object detection with depth-aware transformer,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4012-4021.

[46] H. Zhang, H. Li, X. Liao, F. Li, S. Liu, L. M. Ni, and L. Zhang, ”DA-BEV: Depth aware BEV transformer for 3D object detection,” arXiv e-prints, arXiv-2302, 2023.

[47] H. Song, D. Sun, S. Chun, V. Jampani, D. Han, B. Heo, W. Kim, and M.-H. Yang, ”VIDT: An efficient and effective fully transformer-based object detector,” arXiv preprint arXiv:2110.03921, 2021.

[48] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, ”Multi-class token transformer for weakly supervised semantic segmentation,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 4310-4319.

[49] H. Qiu, Q. Zheng, G. Memmi, J. Lu, M. Qiu, and B. Thuraisingham, ”Deep residual learning-based enhanced JPEG compression in the Internet of Things,” IEEE Transactions on Industrial Informatics, vol. 17, no. 3, pp. 2124-2133, 2020.

[50] H. Qiu, Q. Zheng, T. Zhang, M. Qiu, G. Memmi, and J. Lu, ”Toward secure and efficient deep learning inference in dependable IoT systems,” IEEE Internet of Things Journal, vol. 8, no. 5, pp. 3180-3188, 2020.

[51] Y. Cui, K. Cao, G. Cao, M. Qiu, and T. Wei, ”Client scheduling and resource management for efficient training in heterogeneous IoT-edge federated learning,” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 41, no. 8, pp. 2407-2420, 2021.

[52] K. Gai, K. Xu, Z. Lu, M. Qiu, and L. Zhu, ”Fusion of cognitive wireless networks and edge computing,” IEEE Wireless Communications, vol. 26, no. 3, pp. 69-75, 2019.

[53] C. Li and M. Qiu, Reinforcement Learning for Cyber-Physical Systems: With Cybersecurity Case Studies, Chapman and Hall/CRC, 2019.

[54] Y. Song, Y. Li, L. Jia, and M. Qiu, ”Retraining strategy-based domain adaptation network for intelligent fault diagnosis,” IEEE Transactions on Industrial Informatics, vol. 16, no. 9, pp. 6163-6171, 2019.

[55] Y. Zeng, H. Qiu, G. Memmi, and M. Qiu, ”A data augmentation-based defense method against adversarial attacks in neural networks,” in Algorithms and Architectures for Parallel Processing: 20th International Conference, ICA3PP 2020, New York City, NY, USA, October 2–4, 2020, Proceedings, Part II, pp. 274-289, Springer, 2020.

[56] W. Liang, Y. Yang, C. Yang, Y. Hu, S. Xie, K. Li, and J. Cao, ”PDPChain: A consortium blockchain-based privacy protection scheme for personal data,” IEEE Transactions on Reliability, 2022.

[57] W. Liang, Y. Li, K. Xie, D. Zhang, K. Li, A. Souri, and K. Li, ”Spatial-temporal aware inductive graph neural network for C-ITS data recovery,” IEEE Transactions on Intelligent Transportation Systems, 2022.