

# Image Recognition with Online Lightweight Vision Transformer: A Survey

ZHERUI ZHANG, Beijing University of Posts and Telecommunications, China

RONGTAO XU, State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China

JIE ZHOU, Beijing University of Posts and Telecommunications, China

CHANGWEI WANG, Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology, Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, China

XINGTIAN PEI, Beijing University of Posts and Telecommunications, China

WENHAO XU, Beijing University of Posts and Telecommunications, China

JIGUANG ZHANG, State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China

LI GUO, Beijing University of Posts and Telecommunications, China

LONGXIANG GAO, Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology, Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, China

WENBO XU, Beijing University of Posts and Telecommunications, China

SHIBIAO XU\*, Beijing University of Posts and Telecommunications, China

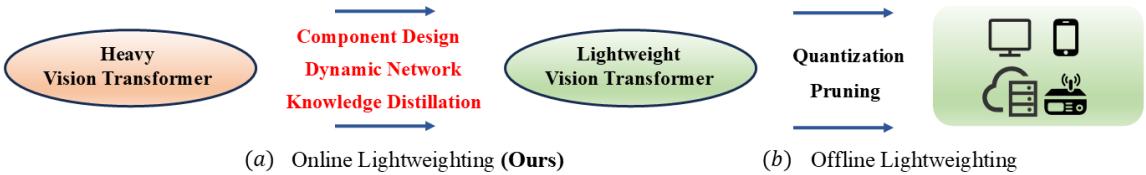
The Transformer architecture has achieved significant success in natural language processing, motivating its adaptation to computer vision tasks. Unlike convolutional neural networks, vision transformers inherently capture long-range dependencies and enable parallel processing, yet lack inductive biases and efficiency benefits, facing significant computational and memory challenges that limit its real-world applicability. This paper surveys various online strategies for generating lightweight vision transformers for image recognition, focusing on three key areas: **Efficient Component Design**, **Dynamic Network**, and **Knowledge Distillation**. We evaluate the relevant exploration for each topic on the ImageNet-1K benchmark, analyzing trade-offs among precision, parameters, throughput, and more to highlight their respective advantages, disadvantages, and flexibility. Finally, we propose future research

---

\*corresponding author

---

Authors' Contact Information: Zherui Zhang, Beijing University of Posts and Telecommunications, Beijing, China, zrz787906410@bupt.edu.cn; Rongtao Xu, State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, xurongtao2019@ia.ac.cn; Jie Zhou, Beijing University of Posts and Telecommunications, Beijing, China, zhoujie8023@bupt.cn; Changwei Wang, Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology, Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China, changweiwang@sdas.org; Xingtian Pei, Beijing University of Posts and Telecommunications, Beijing, China, 2024111242@bupt.cn; Wenhao Xu, Beijing University of Posts and Telecommunications, Beijing, China, xuwenhao@bupt.edu.cn; Jiguang Zhang, State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, jiguang.zhang@ia.ac.cn; Li Guo, Beijing University of Posts and Telecommunications, Beijing, China, guoli@bupt.edu.cn; Longxiang Gao, Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology, Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China, gaolx@sdas.org; Wenbo Xu, Beijing University of Posts and Telecommunications, Beijing, China, xuwb@bupt.edu.cn; Shibiao Xu, Beijing University of Posts and Telecommunications, Beijing, China, shibiaoxu@bupt.edu.cn.



**Fig. 1. ViT Lightweighting Overview.** We categorize existing techniques for alleviating the computational burden of ViT into two steps: online lightweighting and offline lightweighting. (a) Online lightweighting techniques reduce the computational burden during training while ensuring the model structure or expressive capability is kept complete in a soft manner; (b) Offline lightweighting techniques, such as pruning and quantization, trim unnecessary capacity from the model structure. In this paper, we focus more on online lightweighting techniques.

directions and potential challenges in the lightweighting of vision transformers with the aim of inspiring further exploration and providing practical guidance for the community. Project Page: <https://github.com/ajxklo/Lightweight-VIT>

Additional Key Words and Phrases: Model Lightweight, Vision Transformer, Image Recognition, Model Compression, Computer Vision

## 1 Introduction

The Vision Transformer (ViT) [1, 2] rapidly emerge as powerful tools for image recognition, challenging the dominance of the traditional Convolutional Neural Network (CNN) [3–11]. Using the self-attention mechanism of transformers in natural language processing, ViT models encode images as sequences of patches, allowing them to effectively capture global dependencies and contextual information [12], yielding remarkable performance improvements in various vision tasks, including image generation [13–20], object detection [21–31], Embodied Artificial Intelligence [32–43] and semantic segmentation [44–56]. In addition, ViT models show remarkable precision in conventional image recognition benchmarks, achieving near-perfect validation results, such as surpassing an impressive precision 99% on the CIFAR-10 dataset. Consequently, the limitations of small-scale datasets in fully demonstrating the performance superiority of ViT models lead to a shift towards larger benchmark datasets like ImageNet-1K [57] for evaluating recognition tasks.

Despite ViT models impressive performance, they have significant drawbacks in model size and computational overhead. The increased parameters and complex operations slow training and inference and increase resource consumption, limiting practical deployment, especially in resource-constrained environments [58, 59]. To address these issues, a variety of lightweighting techniques are proposed, aiming to reduce the computational burden without substantially compromising model performance.

This survey focuses on online lightweighting strategies[60–62] for ViT models, which differ from offline post-processing methods such as pruning [63–65] and quantization [66–68]. Online techniques aim to optimize the model during training, either by improving the design of internal components [69, 70], transferring knowledge [71–74], or dynamically adjusting the computational path of the network based on the complexity of the input [75, 76]. The goal is to achieve a “soft” lightweighting, maintaining the model architecture and expressive capacity while enhancing efficiency. As illustrated in Figure 1, online lightweighting techniques are seen as a precursor to more aggressive offline post-processing methods.

We categorize online lightweighting techniques for ViT models into three main topics: efficient component design, dynamic network, and knowledge distillation (KD). These topics exploit the ViT property on token-based input to achieve unique efficiency gains. For example, token input updates can be handled using native self-attention, refined linear attention, or even attention-free [77, 78] mechanisms. To address spatial redundancy [79, 80], dynamic token

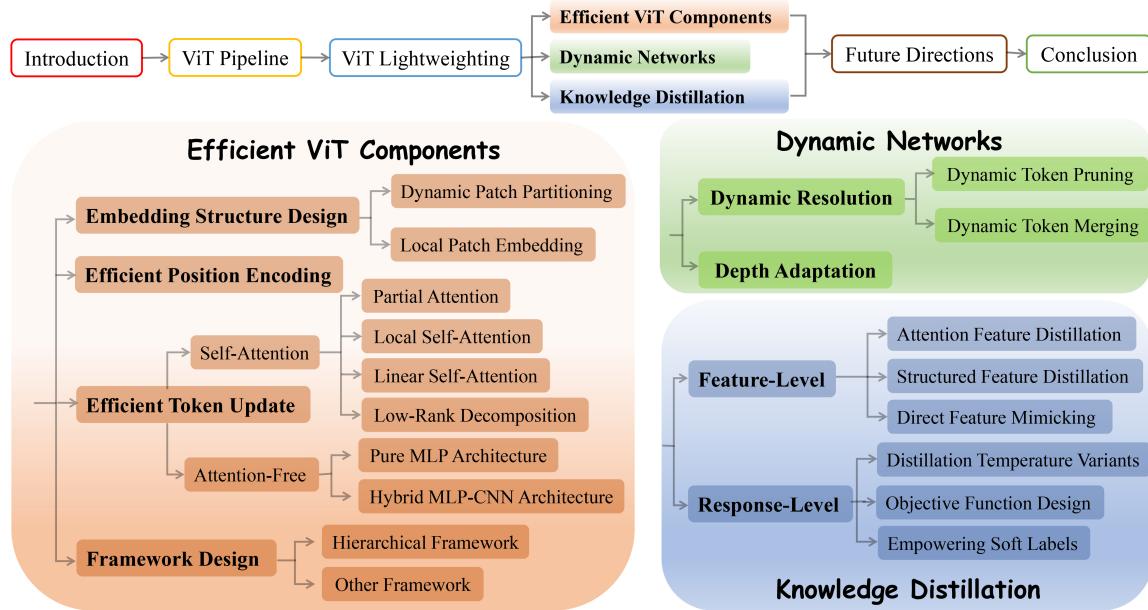


Fig. 2. **Online ViT Lightweighting Overview.** The general overview diagram of this paper shows the basic structure of the paper at the top, and at the bottom lists the methods collated in this paper regarding the online lightweighting of ViT models, mainly efficient ViT components, dynamic networks and knowledge distillation methods.

pruning [75, 81] and token merging [82, 83] strategies are used, offering a more granular approach compared to traditional downsampling of CNN models. Integrating these techniques allows for customized, lightweight strategies developed for real-world applications. Our contributions in this survey are as follows:

- **Comprehensive Overview.** We provide a broad survey of online lightweighting techniques for vision transformer, distinguishing them from offline post-processing techniques, and categorize them into three key topics: efficient component design, dynamic network, and knowledge distillation, with each topic discussed in the context of input flow forwarding.
- **Trade-Off Presentation.** Using ImageNet-1K [57] as the benchmark, we present a detailed analysis of the trade-offs between precision and efficiency for various online lightweighting strategies, offering both qualitative and quantitative insights.
- **Future Directions.** We identify the limitations of current online lightweighting techniques, highlight common assumptions that may need reconsideration, and propose promising research directions, including privacy-sensitive models.

**Difference from Other Surveys** This survey focuses mainly on online lightweighting techniques of ViT models and provides comprehensive experimental reports to uniformly examine the trade-offs existing strategies make between precision and efficiency.

**Survey Overview** This survey uses simplified notation from Table 1. Figure 2 shows the structure: Section 2 introduces the vision transformer concept and pipeline. We then present online lightweight strategies on three main topics. The first topic, discussed in Section 3, explores the **efficient component** design within ViT. The second topic, presented in Section 4, explores **dynamic networks** based on input complexity, including dynamic resolution and depth adaptation.

Table 1. Some Simplified Forms of symbols in this paper.

Usage				
CNN	KD	ViT	Patch or Token	MLP
Convolutional Neural Network	Knowledge Distillation	Vision Transformer	ViT-specific Input Form	Multi-Layer Perceptron
FFN	SA	MHSA	PE	LN
Feed-Forward Network	Self-Attention	Multi-Head Self-Attention	Positional Encoding	Layer Normalization
BN	CE	KL	SOTA	FLOPs
Batch Normalization	Cross Entropy	Kullback-Leibler	State Of The Art	Floating Point Operations per Second
Symbol				
Q	V	K	C	$\Omega$
Query Matrix	Value Matrix	Key Matrix	Embedding Dimension	Computation Complexity

The third topic, in Section 5, focuses on the use of the unique advantages of ViT for **knowledge distillation**. For each topic, we report from a unified perspective on the results of image recognition performance evaluations on the ImageNet [57] benchmark for the methods involved, analyzing their advantages and disadvantages, as well as flexibility. Finally, in Section 6, we offer a discussion on promising directions for future exploration.

## 2 Original ViT Pipeline

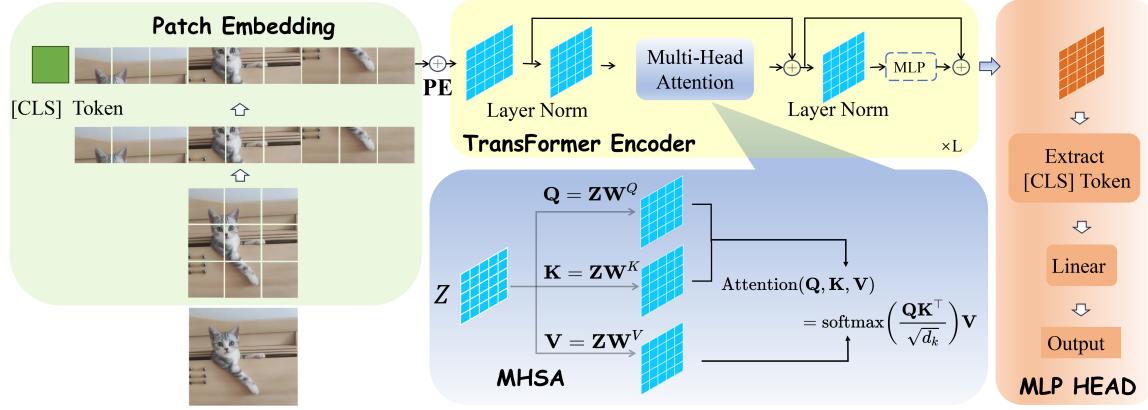
As shown in Figure 3, given an input image  $I$  with resolution  $H \times W$ , the first step in the ViT pipeline involves dividing the image into non-overlapping  $N = (H/P) \times (W/P)$  patches, each of size  $P \times P$ , where  $P$  represents the patch size. Each patch  $\mathbf{x}_p \in \mathbb{R}^{P \times P \times C}$  undergoes a flattening operation to obtain a linear embedding  $\mathbf{x}_p \in \mathbb{R}^{P^2C}$ , followed by a linear projection to obtain the patch embedding  $\mathbf{z}_p^0 = \mathbf{E}\mathbf{x}_p$ , where  $\mathbf{E} \in \mathbb{R}^{D \times P^2C}$  is a learnable projection matrix and  $D$  is the embedding dimension.

To incorporate positional information, a positional embedding  $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$  is added to the patch embeddings. The resulting sequence of embedded patches  $\mathbf{Z}_0 = [\mathbf{z}_{cls}; \mathbf{z}_1^0; \dots; \mathbf{z}_N^0] + \mathbf{E}_{pos}$  serves as the input to the transformer encoder, where  $\mathbf{z}_{cls} \in \mathbb{R}^D$  is a learnable [CLS] token that facilitates recognition.

The transformer encoder consists of  $L$  layers, each comprising a Multi-Head Self-Attention (MHSA) block, a Feed-Forward Network (FFN) or Multilayer Perceptron (MLP). In the MHSA block, the input sequence  $\mathbf{Z}_{l-1}$  is linearly projected to obtain the query  $\mathbf{Q}$ , key  $\mathbf{K}$  and value  $\mathbf{V}$  using the learned projection matrices  $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{D \times D_h}$ , respectively. The attention weights are computed as  $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D_h}})\mathbf{V}$ , where  $D_h$  is the dimension of each head. The outputs of the  $H$  attention heads are concatenated and linearly projected using  $\mathbf{W}_o \in \mathbb{R}^{HD_h \times D}$  to obtain the MSA output.

The FFN consists of two linear transformations with an activation function between, applied positionally to the MHSA output:  $\text{FFN}(\mathbf{Z}) = \text{ReLU}(\mathbf{Z}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$ , where  $\mathbf{W}_1 \in \mathbb{R}^{D \times D_{FFN}}$ ,  $\mathbf{b}_1 \in \mathbb{R}^{D_{FFN}}$ ,  $\mathbf{W}_2 \in \mathbb{R}^{D_{FFN} \times D}$ , and  $\mathbf{b}_2 \in \mathbb{R}^D$  are learnable parameters. Each transformer encoder layer incorporates residual connections and layer normalization (LN), resulting in the output:  $\mathbf{Z}_l = \text{LN}(\mathbf{Z}_{l-1} + \text{FFN}(\text{LN}(\mathbf{Z}_{l-1} + \text{MSA}(\mathbf{Z}_{l-1}))))$ .

After passing through the  $L$  transformer encoder layers, the final output sequence  $\mathbf{Z}_L = [\mathbf{z}_{cls}^L; \mathbf{z}_1^L; \dots; \mathbf{z}_N^L]$  is obtained. The output of the [CLS] token  $\mathbf{z}_{cls}^L$  is used for recognition by applying a linear transformation followed by softmax activation:  $\mathbf{y} = \text{softmax}(\mathbf{W}_{cls}\mathbf{z}_{cls}^L + \mathbf{b}_{cls})$ , where  $\mathbf{W}_{cls} \in \mathbb{R}^{D \times K}$ ,  $\mathbf{b}_{cls} \in \mathbb{R}^K$ , and  $K$  is the number of classes.



**Fig. 3. Original ViT Overview.** The pipeline of the original ViT for image recognition is divided into four main components. Firstly, Patch Embedding is used to obtain input in token form. Secondly, there is a stack of Transformer layers, which includes the Multi-Head Self-Attention (MHSA) mechanism, facilitating ViT global comprehension ability. Finally, a Multi-Layer Perceptron (MLP) structure is employed to project into category space.

### 3 Efficient ViT Components

As the desire for improved image recognition precision continues to grow, challenges arise from image quality, object characteristics, and potential dependencies. The vision transformer stands out as particularly well suited for such tasks because of the unique representation advantages and robust adaptability to global understanding. Consequently, novel designs for various components of ViT emerge in abundance, accompanied by additional training and inference costs.

In this section, we present the efforts towards online lightweight design of each component in the order in which it processes the input image within the network. This includes **Embedding Structure Design** in Section 3.1, **Efficient Position Encoding** in Section 3.2, **Efficient Token Update** in Section 3.3, and **Framework Design** in Section 3.4.

#### 3.1 Embedding Structure Design

Patch embedding serves as the initial stage in the ViT pipeline, responsible for converting the input image into a sequence of tokens that can be processed by the transformer encoder. This stage plays a crucial role in determining the efficiency of the overall model, as the number of patches directly influences the computational complexity of subsequent self-attention layers. Specifically, we discuss two prominent directions: **Dynamic Patch Partitioning (DPP)** and **Local Patch Embedding (LPE)**. DPP explores adaptive partitioning of the input image to generate effective patches at multiple scales and multiple views. On the other hand, LPE leverages CNN models to extract rich local information.

**3.1.1 Dynamic Patch Partitioning.** DPP adapts to the computational demands by employing patches of varying sizes or dynamically adjusting patch dimensions. CrossViT<sup>1</sup> [84] employs a dual-branch Transformer with multi-scale patches, while SSA [88] merges tokens to represent larger objects efficiently. TopFormer<sup>2</sup> [91] integrates multi-scale features using token and semantic pyramid modules, and T2T-ViT<sup>3</sup> [92] adopts a hierarchical approach to model local structures. Dynamic patch partitioning techniques, such as IA-RED<sup>2</sup> [81], DynamicViT [93], and [89], eliminate redundant or non-essential patches to optimize computational efficiency.

<sup>1</sup><https://github.com/IBM/CrossViT>

<sup>2</sup><https://github.com/hustvl/TopFormer>

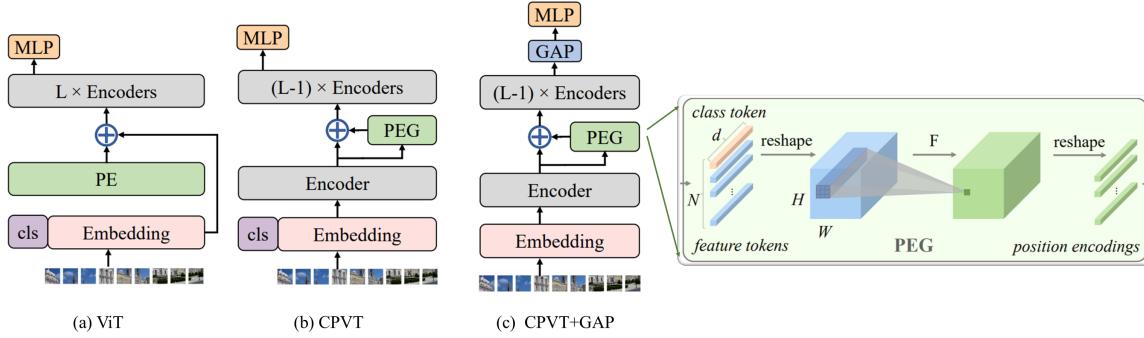
<sup>3</sup><https://github.com/yitu-opensource/T2T-ViT>

Table 2. A quantitative comparative analysis is conducted on the classic methods involved in **Embedding Structure Design** (Section 3.1), using ImageNet-1K [57] as the benchmark, with the resolution standardized to  $224 \times 224$ .  $\dagger$  denotes that the CrossViT model is modified by replacing the linear patch embedding in ViT with three convolutional layers as patch labelers. The best-performing method is highlighted in bold, and the runner-up is underlined.

Model	Year	Params (M) $\downarrow$	FLOPs (G) $\downarrow$	Top-1 Acc. (%) $\uparrow$	Model	Year	Params (M) $\downarrow$	FLOPs (G) $\downarrow$	Top-1 Acc. (%) $\uparrow$
▲ CrossViT-Ti DeiT-Ti [84]	2021	6.9	1.6	73.4	● Mobile-Former-96M [87]	2022	4.6	-	72.8
▲ CrossViT-9 DeiT-Ti [84]	2021	8.6	1.8	73.9	● Mobile-Former-151M [87]	2022	7.6	-	75.2
▲ CrossViT-9 $\dagger$  DeiT-Ti [84]	2021	8.8	2.0	77.1	● Mobile-Former-214M [87]	2022	9.4	-	76.7
▲ CrossViT-S DeiT-S [84]	2021	26.7	5.6	81.0	● Mobile-Former-294M [87]	2022	11.4	-	77.9
▲ CrossViT-15 DeiT-S [84]	2021	27.4	5.8	81.5	● Mobile-Former-508M [87]	2022	14.0	-	79.3
▲ CrossViT-15 $\dagger$  DeiT-S [84]	2021	28.2	6.1	82.3	● SSA-T [88]	2022	11.3	2.1	79.8
▲ CrossViT-18 DeiT-B [84]	2021	43.3	9.0	82.5	● SSA-S [88]	2022	22.4	4.9	82.9
▲ CrossViT-18 $\dagger$  DeiT-B [84]	2021	44.3	9.5	82.8	● SSA-B [88]	2022	39.6	8.1	84.0
▲ CrossViT-B DeiT-B [84]	2021	104.7	21.2	82.2	● SSA-L [88]	2022	81.2	14.9	84.6
▲ IA-RED <sup>2</sup>  DeiT-S [81]	2021	22.1	-	79.1	■ DPS-ViT DeiT-Ti [89]	2022	5.7	0.6	72.1
▲ IA-RED <sup>2</sup>  DeiT-B [81]	2021	86.6	11.8	80.3	■ DPS-ViT T2T-ViT [89]	2022	21.5	3.1	81.3
▲ CvT-13 [85]	2021	20.0	4.5	81.6	■ DPS-ViT DeiT-S [89]	2022	22.1	2.6	79.4
▲ CvT-21 [85]	2021	32.0	7.1	82.5	■ DPS-ViT DeiT-B [89]	2022	86.6	9.4	81.6
▲ MobileViT-XS [86]	2022	2.3	0.7	74.8	■ RepViT-M0.9 [90]	2024	5.4	1.6	78.7
▲ MobileViT-S [86]	2022	5.6	-	78.4	■ RepViT-M1.0 [90]	2024	6.8	2.2	80.0
● Mobile-Former-26M [87]	2022	3.2	-	60.4	■ RepViT-M1.1 [90]	2024	8.2	2.6	80.7
● Mobile-Former-52M [87]	2022	3.5	-	68.7	■ RepViT-M2.3 [90]	2024	22.9	9	83.3

**3.1.2 Local Patch Embedding.** LPE incorporates convolutional layers into the patch embedding process, leveraging its local feature extraction capabilities while preserving the ViT global context understanding. MobileViT [86] uses convolutional layers to encode pixel information within non-overlapping patches. MobileFormer [87] integrates a lightweight MobileNet, using global tokens to fuse local and global features. CvT [85] employs a hierarchical design with convolutional token embedding layers at each stage.

**3.1.3 Summary: Embedding Structure Design .** The traditional global patch encoding strategy does not adequately capture local information from images, affecting the precision of fine-grained image recognition on ImageNet [57], while a dense patch encoding approach incurs a higher computational burden. Two directions for improvement under this topic involve multi-scale and multi-perspective adaptive partitioning of input images, combined with CNNs to extract rich local features, enhancing the quality and efficiency of patch embedding. In Table 2, we report the relevant experimental results, where the following phenomena can be observed: **(i)** The CrossViT [84] series introduces dynamic patch partitioning and local partition embedding mechanisms, significantly outperforming the traditional DeiT model, with competitive FLOPs and an accuracy improvement from 81.0%  $\rightarrow$  82.3%; **(ii)** Lightweight variants such as MobileViT [86] and CvT [85] achieve an excellent balance between accuracy and efficiency, for instance, MobileViT-S [86] achieves an accuracy of 78.4% with just 5.6M parameters; **(iii)** State-of-the-art (SOTA) methods like SSA [88] and DPS-ViT [89] achieve a win-win in both accuracy and efficiency, with SSA-B [88] achieving a high accuracy of 84.0% with only 39.6M parameters.



**Fig. 4. How Positional Encoding Works.** A comparative illustration [94] of the vision transformer and its modified counterpart is presented. (a) depicts the ViT model that employs explicit learnable positional encodings. In contrast, (b) showcases the Conditional Positional Vision Transformer (CPVT) utilizing conditional positional encoding generated by the proposed Position Encoding Generator (PEG) plugin, which serves as the default configuration. (c) illustrates the CPVT-GAP variant, which, in the absence of a [CLS] token, performs global average pooling (GAP) across all sequence elements.

### 3.2 Efficient Position Encoding

After converting the image input into a sequence of patches, it is necessary to encode the spatial information lost during patch partitioning. This is where Position Encoding (PE) comes into play, as shown in Figure 4 (a). Traditional absolute positional encodings, such as the widely used sinusoidal method:

$$\text{PE} = \begin{cases} \sin\left(\frac{\text{pos}}{10000^{2k/d}}\right) & \text{if } i = 2k \\ \cos\left(\frac{\text{pos}}{10000^{2k/d}}\right) & \text{if } i = 2k + 1 \end{cases} \quad (1)$$

where  $k = 1, \dots, d/2$ ,  $i$  and  $d$  denote the index and length of the vector respectively, and “pos” denotes the position of the element in the sequence. While simple, this method suffers from limitations such as lack of adaptability to variable input sizes or non-available prior information. Recent research has shifted towards more efficient and flexible position encoding techniques that can effectively capture spatial relationships while minimizing computational overhead.

One promising direction is to incorporate relative positional information directly into the attention mechanism. RoPE [95] achieves this by integrating query and key vectors, improving the ability of self-attention to encode spatial relationships. Similarly, PoSGU [96] embeds positional information within the token mixing layer, maintaining expressiveness while reducing parameter overhead. Another approach leverages efficient computation methods for relative positional embeddings. LISA [97] employs the Fast Fourier Transform (FFT) to approximate relative positional embeddings, achieving log-linear complexity. FIRE [98] introduces gradient interpolation techniques for relative encoding, and LRPE [99] proposes efficient linear complexity encodings.

Furthermore, adapting positional encoding to varying input contexts shows significant promise. For example, as shown in Figure 4, CPVT [94] dynamically incorporates positional information by conditioning image patch content, thus enhancing model flexibility and performance. This is particularly beneficial for tasks such as extrapolating lengths in transformers, significantly improving medical segmentation tasks with Polyp-ViT [100].

**3.2.1 Summary: Efficient Position Encoding.** ViT is highly dependent on the self-attention mechanism for outstanding global understanding, which inherently lacks the ability to perceive the sequential or spatial relationships of the input data. Effectively encoding positional information can enhance performance in several aspects: (i). Facilitates

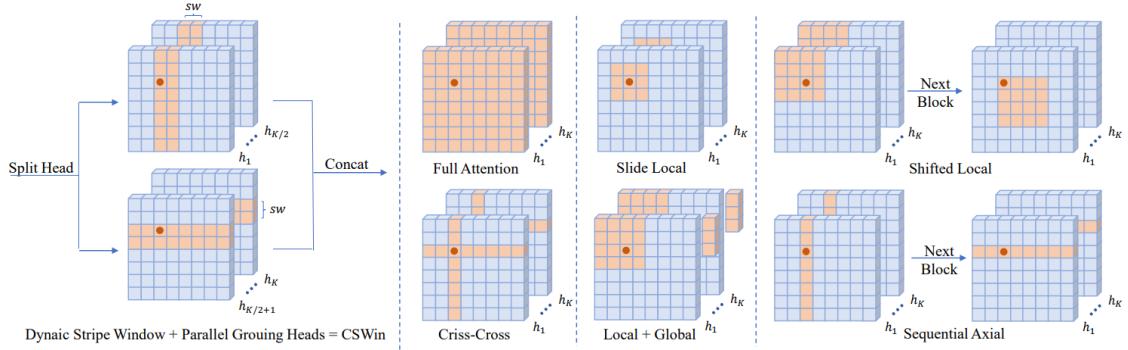


Fig. 5. **Improvements in Self-Attention.** CSwin [101] splits multi-heads ( $h_1, \dots, h_K$ ) into two groups and performs self-attention in horizontal and vertical stripes simultaneously. Then, CSwin adjusts the stripe width according to the network depth, which achieves a better trade-off between computational cost and capability.

the ViT model in comprehending the spatial structure of images, enabling it to differentiate objects and their positions within the frame, which is particularly crucial for ImageNet [57], which contains a vast number of objects with diverse spatial configurations; **(ii).** Allowing the ViT model to focus more effectively on relevant regions of the image; **(iii).** Eliminates ambiguity in feature representations by ensuring that the position of each pixel or patch is known; **(iv).** Accelerates the ViT model convergence speed during the training process by providing clear spatial signals, making the learning of large-scale image recognition datasets more reliable.

### 3.3 Efficient Token Update

**3.3.1 Self-Attention.** With the spatial context re-introduced, the core component of the ViT architecture comes into play: Self-Attention (SA). This mechanism empowers the model to capture long-range dependencies among patch inputs, enabling the comprehensive understanding of the image. However, the computational complexity of standard self-attention grows quadratically with the number of tokens, posing a significant bottleneck for resource-constrained applications.

To mitigate this issue, researchers have made significant efforts to develop efficient self-attention mechanisms that maintain the ability of the model to capture global dependencies while reducing computational cost. This section explores four prominent directions in efficient self-attention design: **Partial Attention**, **Local Self-Attention**, **Linear Self-Attention**, and **Low-rank Decomposition**.

**Partial Attention (PA).** PA focuses on limiting the scope of attention to a subset of tokens, thereby reducing computational demands. The Swin Transformer<sup>4</sup> [110] achieves this by using shifted windows to confine self-attention computations to non-overlapping local windows, resulting in linear computational growth. Specifically, assuming each window contains  $M \times M$  pixels, the computational complexities of the standard global Multi-Head Self-Attention (MHSA) module and the Window-based Multi-head Self-Attention (WMSA) on the input image with  $h \times w$  patches are as follows:

$$\begin{aligned}\Omega(\text{MSA}) &= 4hwC^2 + 2(hw)^2C \\ \Omega(\text{WMSA}) &= 4hwC^2 + 2M^2hwC\end{aligned}\tag{2}$$

<sup>4</sup><https://github.com/microsoft/Swin-Transformer>

Table 3. A quantitative comparative analysis is conducted on the classic methods involved in **Self-Attention Optimization** (Section 3.3.1), using ImageNet-1K [57] as the benchmark, with the resolution standardized to  $224 \times 224$ . The best-performing method is highlighted in bold, and the runner-up is underlined.

Model	Year	Params (M)↓	FLOPs (G)↓	Top-1 Acc. (%)↑
● GFNet-Ti [102]	2021	7.0	1.3	74.6
● GFNet-XS [102]	2021	16.0	2.9	78.6
● GFNet-S [102]	2021	25.0	4.5	80.0
● GFNet-B [102]	2021	43.0	7.9	80.7
● Focal-Tiny [103]	2021	29.1	4.9	82.2
● Focal-Small [103]	2021	51.1	9.1	83.5
● Focal-Base [103]	2021	89.8	16.0	83.8
● T2T-ViT-14 [92]	2021	21.5	4.8	81.5
● T2T-ViT-19 [92]	2021	39.2	8.5	81.9
● T2T-ViT-24 [92]	2021	64.1	13.8	82.3
▲ CSWin-T [101]	2022	23.0	4.3	82.7
▲ CSWin-S [101]	2022	35.0	6.9	83.6
▲ CSWin-B [101]	2022	78.0	15.0	84.2
▲ MaxSA-T [104]	2022	31.0	5.6	83.6
▲ MaxSA-S [104]	2022	69.0	11.7	84.5
▲ MaxSA-B [104]	2022	120.0	23.4	85.0
▲ MaxSA-L [104]	2022	212.0	43.9	85.2
■ SKIPAT ViT-T [105]	2023	5.8	1.1	72.9
■ SKIPAT ViT-S [105]	2023	22.1	4.0	80.2
■ SKIPAT-B ViT-B [105]	2023	86.7	15.2	82.2
■ SwiftFormer-XS [106]	2023	3.5	-	75.7
■ SwiftFormer-S [106]	2023	6.1	-	78.5
■ SwiftFormer-L1 [106]	2023	12.1	-	80.9
■ BiFormer-T [107]	2023	13.1	2.2	81.4
■ BiFormer-S [107]	2023	26.0	4.5	83.8
■ BiFormer-B [107]	2023	57.0	9.8	84.3
■ REMViTv2-T [108]	2024	24.0	4.7	82.7
■ REViT-T [108]	2024	29.0	4.5	81.5
■ RESwim-T [108]	2024	29.0	4.5	81.5
■ REViT-B [108]	2024	86.0	17.5	82.4
■ SSViT-T [109]	2024	15.0	1.2	83.0
■ SSViT-S [109]	2024	27.0	2.2	84.4
■ SSViT-B [109]	2024	57.0	4.8	85.3
■ SSViT-L [109]	2024	100.0	9.1	85.7

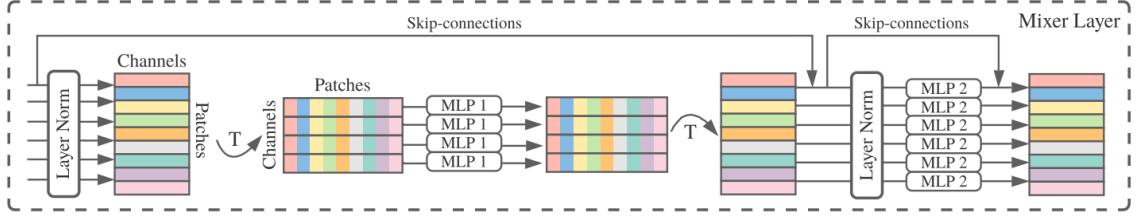
where the C denotes the embedding dimension, symbol  $\Omega$  is used to measure the computation complexity,  $\Omega(\text{MHSA})$  has a quadratic complexity with respect to the number of patches  $h \times w$ , while the  $\Omega(\text{WMSA})$  has a promising linear complexity. Similarly, the CSWin Transformer<sup>5</sup> [101] partitions the input features into horizontal and vertical stripes and computes SA within these localized regions, as shown in Figure 5. The TNT<sup>6</sup> model [111] employs transformers at the patch and pixel levels, facilitating detailed attention to local features.

**Local Self-Attention.** Local SA takes the idea of partial attention by restricting attention operations to local regions of the input image [112]. Multi-axis attention (Max-SA) [104] simplifies the complexity of attention from quadratic to linear by implementing windowed and mesh attention, while ReViT [108] addresses feature collapse in deep ViT layers through residual attention learning. Deformable self-attention [113] and dynamic sparse attention [107, 114] dynamically select attention blocks based on data dependencies and query patterns, respectively. The Focal Self-attention model [103] captures both short-range and long-range dependencies, utilizing fine-grained local attention and coarse-grained distant interactions. ViL [115] combines multi-scale architectures with extended context considerations, improving high-resolution image processing and encoding.

**Linear Self-Attention.** Linear SA aims to reduce the computational complexity of SA from quadratic to linear [117]. Various methods are proposed, such as replacing dot product attention with position-sensitive hashing [118]. GFNet [102] employs discrete two-dimensional Fourier transforms combined with a learnable global filter, and Linformer [119] introduces a low-rank matrix to achieve linear complexity. SwiftFormer [106] uses an efficient additive attention

<sup>5</sup><https://github.com/microsoft/CSWin-Transformer>

<sup>6</sup><https://github.com/lucidrains/transformer-in-transformer>



**Fig. 6. Attention-Free Paradigm for Token Updates.** MLP-Mixer [116] proposes to replace convolution and attention in ViT with MLP models. The figure shows the key part of the MLP-Mixer, which contains mixing patches and mixing channels.

mechanism, while [120] proposes separable self-attention using element-wise intelligent operations. The neighborhood attention mechanism [121] adjusts attention weights based on the 2D Manhattan distance from neighboring patches. Similarly, Vision Outlooker (VOLO) [122], while primarily focusing on computer vision tasks, introduces an innovative locality-sensitive attention mechanism, balancing efficiency and accuracy.

**Low-Rank Decomposition.** These techniques decompose the attention matrix into low-dimensional representations to optimize the computation. MobileViT [123] combines low-rank decomposition and sparsification techniques to transform ViT into a mobile-friendly version. The Low-Rank Transformer [124] reduces computational complexity by decomposing the SA layer, while the Performer [125] integrates kernel techniques with low-rank decomposition. Tensor decomposition techniques, such as CP-Decomposition [126, 127], and factorized self-attention mechanisms [128] have also been applied for model optimization.

**3.3.2 Attention-Free Architecture.** While self-attention revolutionizes the field of image recognition, its computational complexity motivates the exploration of alternative architectures that achieve comparable performance with reduced resource requirements and leads to renewed interest in MLP, a classic neural network architecture consisting of fully connected layers and activation functions.

MLP-based architectures, often dispensing with the complexities of convolutional and self-attention mechanisms, offer advantages in terms of adaptability and computational efficiency. This topic focuses on two main directions in MLP research for efficient image recognition: **Pure MLP Architecture** and **Hybrid MLP-CNN Architecture**. These approaches highlight the potential of the MLP design as a competitive alternative to attention-based models.

**Pure MLP Architecture.** Pure MLP architectures aim to simplify the token update operation for improved efficiency while maintaining performance. ResMLP [129] introduces a residual network architecture that alternates between linear layers and two-layer feedforward networks. As shown in Figure 6, MLP-Mixer<sup>7</sup> [116] introduces Token-Mixing and Channel-Mixing MLP blocks to process images, avoiding convolutions and SA. S2-MLP [130], inspired by MLP-Mixer, relies on channel mixing and incorporates a spatial shift operation for patch interaction. AS-MLP<sup>8</sup> [131] adopts axially shifted channels to capture local dependencies, allowing the pure MLP model to simulate the local receptive fields of CNN models. MLP-Unet [132] applies a fully MLP-based method and cascaded upsampler to keep local information effectively.

**Hybrid MLP-CNN Structure.** Hybrid attention-free structures that integrate MLP and CNN offer a promising balance of complementary strengths. X-MLP [142] employs linear projections for feature map interactions, eliminating the

<sup>7</sup><https://github.com/jaketae/mlp-mixer>

<sup>8</sup><https://github.com/svip-lab/AS-MLP>

Table 4. A quantitative comparative analysis is conducted on the classic methods involved in **Attention-Free Design** (Section 3.3.2), using ImageNet-1K as the benchmark, with the resolution standardized to  $224 \times 224$ . The “/16” in Mixer means the model of base scale with patches of resolution  $16 \times 16$ . The best-performing method is highlighted in bold, and the runner-up is underlined.

Model	Year	Params (M)↓	FLOPs (G)↓	Top-1 Acc. (%)↑	Model	Year	Params (M)↓	FLOPs (G)↓	Top-1 Acc. (%)↑
▲ GMLP-T [133]	2021	6.0	-	72.3	■ CycleMLP-B1 [137]	2022	15.0	2.1	79.1
▲ GMLP-S [133]	2021	20.0	-	79.6	■ CycleMLP-B2 [137]	2022	27.0	3.9	81.6
▲ GMLP-B [133]	2021	73.0	-	81.6	■ CycleMLP-T [137]	2022	28.0	4.4	81.3
▲ HireMLP-Tiny [134]	2021	18.0	2.1	79.7	■ CycleMLP-B3 [137]	2022	38.0	6.9	82.4
▲ HireMLP-Small [134]	2021	33.0	4.2	82.1	■ CycleMLP-S [137]	2022	50.0	8.5	82.9
▲ HireMLP-Base [134]	2021	58.0	8.1	83.2	■ CycleMLP-B [137]	2022	88.0	15.2	83.4
▲ HireMLP-Large [134]	2021	96.0	13.4	83.8	■ S <sup>2</sup> -MLP-deep [138]	2022	51.0	10.5	80.7
▲ Mixer-B/16 [116]	2021	59.0	11.6	82.4	■ S <sup>2</sup> -MLP-wide [138]	2022	71.0	14.0	80.0
▲ Mixer-L/16 [116]	2021	207.0	-	77.1	■ PoolFormer-S12 [139]	2022	12.0	1.8	77.2
▲ ConvMLP-S [135]	2021	9.0	2.4	76.8	■ PoolFormer-S24 [139]	2022	21.0	3.4	80.3
▲ ConvMLP-M [135]	2021	17.4	3.9	79.0	■ PoolFormer-S36 [139]	2022	31.0	5.0	81.4
▲ ConvMLP-L [135]	2021	42.7	9.9	80.2	■ PoolFormer-M36 [139]	2022	56.0	8.8	82.1
■ AS-MLP-T [131]	2022	28.0	4.4	81.3	■ PoolFormer-M48 [139]	2022	73.0	-	82.5
■ AS-MLP-S [131]	2022	50.0	8.5	83.1	● SVT-H-S [140]	2023	22.0	3.9	84.2
■ AS-MLP-B [131]	2022	88.0	15.2	83.3	● SVT-H-B [140]	2023	32.8	6.3	<u>85.2</u>
■ ResMLP-S12 [136]	2022	15.0	3.0	76.6	● SVT-H-L [140]	2023	54.0	12.7	<b>85.7</b>
■ ResMLP-S24 [136]	2022	30.0	6.0	79.4	■ MONet-T [141]	2024	10.3	2.8	77.0
■ ResMLP-B24 [136]	2022	116.0	23.0	81.0	■ MONet-S [141]	2024	32.9	6.8	81.3

need for convolution operations in patch embedding. [143] propose a pre-training method that does not rely on attention mechanisms. gMLP [133] identifies the high computational burden of self-attention and uses channel and static spatial projections to maintain high precision with fewer parameters. A hybrid model [144] achieves performance comparable to CNN-Transformer [145] hybrids by employing matrix decomposition and multiscale information fusion. [146] propose a structured state space method for efficiently modeling long sequences. US-MLP [147] integrates CNN layers with sparse MLP blocks. Furthermore, GSS model [148] effectively processes long sequence modeling through a gated state space, providing immense potential for image recognition applications. Similarly, [146] addresses the computational bottleneck in long sequence processing by optimizing state space modeling. PoolFormer [139]<sup>9</sup> replaces self-attention with grouping operations as a token mixer. Finally, Mamba [149] achieves linear time complexity in sequence modeling through selective state space. By applying a selective state space, Mamba effectively balances accuracy and computational burden in sequence modeling [150–154].

**3.3.3 Summary: Efficient Token Update. Self-Attention Optimization.** While self-attention captures long-distance dependencies, its computational complexity increases with large datasets like ImageNet-1K. The goal is to optimize complexity without losing global dependency capture. We report the relevant experimental results in Table 3, where it can be observed that: *(i)*. Methods based on local self-attention, such as GFNet [155] and SKIPAT [105], slightly compromise accuracy but significantly reduce computational burden. For example, SKIPAT [105] reduces the parameters to 5.8M and the FLOPs to 1.1G; *(ii)*. BiFormer [107] achieves an excellent balance between parameters, computational cost, and accuracy, with 13.1M parameters and only 2.2G FLOPs, while attaining an accuracy of 81.4%; *(iii)*. SSViT [109],

<sup>9</sup><https://github.com/sail-sg/poolformer>

Table 5. A quantitative comparative analysis is conducted on the classic methods involved in **Efficient Framework Design** (Section 3.4), using ImageNet-1K as the benchmark, with the resolution standardized to  $224 \times 224$ . The best-performing method is highlighted in bold, and the runner-up is underlined.

Model	Year	Params (M)↓	FLOPs (G)↓	Top-1 Acc. (%)↑
▲ PVT-Tiny [156]	2021	13.2	1.9	75.1
▲ PVT-Small [156]	2021	24.5	3.8	79.8
▲ PVT-Medium [156]	2021	44.2	6.7	81.2
▲ PVT-Large [156]	2021	61.4	9.8	81.7
▲ HAT-Net-Tiny [157]	2021	12.7	2.0	79.8
▲ HAT-Net-Small [157]	2021	25.7	4.3	82.6
▲ HAT-Net-Medium [157]	2021	42.9	8.3	84.0
▲ HAT-Net-Large [157]	2021	63.1	11.5	84.2
▲ Twims-SVT-S [158]	2021	24.0	2.9	81.7
▲ Twims-SVT-S [158]	2021	24.1	3.8	81.2
▲ Twims-SVT-B [158]	2021	56.0	8.6	83.2
▲ Twims-PCPVT-B [158]	2021	43.8	6.7	82.7
▲ Twims-PCPVT-L [158]	2021	60.9	9.8	83.1
▲ Twims-SVT-L [158]	2021	99.2	15.1	83.7
▲ Swim-T [110]	2021	29.0	4.5	81.3
▲ Swim-S [110]	2021	50.0	8.7	83.0
▲ Swim-B [110]	2021	88.0	15.4	83.5
● MaxViT-T [104]	2022	31.0	5.6	83.6
● MaxViT-S [104]	2022	69.0	11.7	84.5
● MaxViT-B [104]	2022	120.0	23.4	85.0
● MaxViT-L [104]	2022	212.0	43.9	85.2
■ FDViT-Ti [159]	2023	4.5	0.6	73.7
■ FDViT-S [160]	2023	21.5	2.8	81.5
■ FDViT-B [160]	2023	67.8	11.9	82.4
■ HiViT-T [161]	2023	19.2	4.6	82.1
■ HiViT-S [161]	2023	38.5	9.1	83.5
■ HiViT-L [161]	2023	66.4	15.9	83.8
■ Hiera-T [162]	2023	28.0	5.0	82.8
■ Hiera-S [162]	2023	35.0	6.0	83.8
■ Hiera-B [162]	2023	52.0	9.0	84.5
■ Hiera-B+ [162]	2023	70.0	13.0	85.2
■ Hiera-L [162]	2023	214.0	40.0	86.1
■ Hiera-H [162]	2023	673.0	125.0	86.9
■ HIRI-ViT-S [163]	2024	34.8	4.5	83.6
■ HIRI-ViT-B [163]	2024	54.4	8.2	84.7
■ HIRI-ViT-L [163]	2024	94.4	17.0	85.3

as a state-of-the-art method, demonstrates an impressive accuracy of 85.7% with only 9.1G FLOPs, which is lower than the DeiT standard model, indicating that optimization of self-attention still holds potential.

**Attention-Free Design.** The MLP design enhances computational efficiency by eliminating complex attention calculations, enabling faster adaptation to various hardware and optimization strategies. When combined with CNNs, MLPs leverage their computational efficiency while keeping the exceptional feature extraction capabilities of CNNs, achieving a balance between performance and efficiency. Table Y presents comparative results of relevant methods, revealing the following observations: *(i)*. HireMLP [134] series exhibit outstanding performance with relatively fewer parameters and lower FLOPs. For example, HireMLP-Small achieves an accuracy of 82.1% using only 33M parameters and 4.2G FLOPs. *(ii)*. The hybrid MLP-CNN architecture can be adjusted according to specific requirements, accommodating various application scenarios and resource constraints. HireMLP-Small [134] achieves an accuracy of 82.1% with only 4.2 G FLOPs. *(iii)*. Particularly notable is SVT-H-L [121], which achieves an impressive Top-1 accuracy of 85.7% using only 54M parameters and 12.7G FLOPs.

### 3.4 Framework Design

Inspired by hierarchical structures in convolutional neural networks for feature extraction and image pyramiding, researchers are integrating these principles into vision transformer (ViT) architectures. This hierarchical design aims to improve the efficiency of ViT and the learning of representation. Hierarchical ViTs differentiate background and foreground objects, improving image recognition in complex scenes with varied scales or occlusions [164]. This topic

explores hierarchical ViT design, with the aim of capturing features at multiple scales for both global context and fine details.

**Hierarchical Framework.** Swin Transformer [110] leads the integration of hierarchical structure into ViT models, introducing shifted window attention and subsampling techniques. PVT [156] and PVTv2 [96] address these issues using a pyramid structure, linear complexity attention, and overlapping patch embeddings. FDViT [159] introduces an innovative downsampling module to address the issue of excessive information loss caused by traditional integer-stride downsampling, and allows the reduction of feature map dimensions with non-integer strides, thereby smoothly decreasing computational complexity while preserving more information. The Twins [158] architecture introduces two hierarchical ViT variants: Twins-PCPVT and Twins-SVT. Twins-PCPVT improves the performance of PVT [156] through conditional positional encoding, and Twins-SVT employs a robust local-global decoupled SA mechanism, enabling exceptional local detail capture and global information processing. MaxViT [104] introduces multi-axis attention blocks (Max-SA) to blend local and global paradigms, maintaining efficacy at different resolutions. H-MHSA [165] further refines these approaches by maintaining granularity while modeling global and local dependencies. CvT<sup>10</sup> [85] takes a different route by integrating convolutions directly into the vision transformer architecture, using the hierarchical feature extraction capabilities of the CNN models to create a multi-stage structure. This approach bridges the gap between the CNN and ViT models, combining their strengths. Similarly, PiT<sup>11</sup> [166] proposes methods to better handle spatial dimensions, improving the hierarchical processing of spatial information.

**Other Framework.** For the efficiency exploration aspect of the novelty structure, Mask R-CNN [167] demonstrates how strategic partitioning can enhance task performance through various upsampling or downsampling operations. ViT-Adapter [168] introduces a pre-training-free adapter for more effective inductive bias transfer. Using self-supervised techniques, MixMAE [169] by using visible tokens from different images for reconstruction reduces computational demand. HiViT [170] improves efficiency by excluding all masked tokens. Complementary approaches, such as the asymmetric encoder-decoder structure [171], significantly reduce memory usage by clustering attention and sparse convolutions. Hiera [162] advances these ideas by removing non-transformer elements and enhancing image recognition analysis.

**3.4.1 Summary: Framework Design.** Traditional ViTs generate single-resolution feature maps, limiting multi-scale information capture for image recognition. By introducing a hierarchical structure design into the ViT architecture, it gains the ability to extract fine-grained features and effectively overcomes isotropy, enabling it to adapt to a diverse range of recognition tasks. We provide relevant experimental results in Table 5, where it can be observed that models utilizing hierarchical structures for multi-scale feature extraction generally exhibit superior performance in terms of accuracy. In particular, the SOTA Hiera-H [162] not only possesses exceptional parameters and FLOPs, but also achieves the highest accuracy of 86.9%. HAT-Net and Hiera offer diverse optimizations in parameters and computational complexity compared to traditional ViTs and can be flexibly adapted to models of varying capacities to accommodate practical application requirements.

### 3.5 Efficient ViT Component Discussion

In this section, we focus on the internal components of the vision transformer and the efforts made to achieve a balance between efficiency and precision. We present the methods in the order of the image input flow through the network,

<sup>10</sup><https://github.com/microsoft/CvT>

<sup>11</sup><https://github.com/naver-ai/pit>

specifically highlighting four aspects of technical innovation: Embedding Structure Design, Efficient Position Encoding, Efficient Token Update, and Framework Design. In addition, we discuss the relevance of these components in the broader scope of image recognition, emphasizing how they contribute to the overall accuracy and speed of the model.

Tables 2, 3, 4 and 5 present the precision benchmarks of these methods and its variants on the ImageNet dataset, highlighting the trade-offs achieved between performance and efficiency at a consistent resolution of  $224 \times 224$ . From these tables, several conclusions can be drawn:

- a. Advancements in patch embedding and positional encoding, often implemented independently, flexibly integrate into existing frameworks, mitigating the need for extensive hyperparameter tuning and specialized optimization techniques;
- b. Locality emerges as a driving force in enhancing ViT component performance. This is evident in patch embedding that focuses on local regions through the attention mechanism, the reduction of self-attention scope or integration of CNN models, and the hierarchical information extraction in framework design. These strategies yield significant efficiency gains without compromising performance;
- c. While shifting from self-attention to attention-free designs demonstrably reduces FLOPs, recent self-attention optimization designs demonstrate a compelling balance between performance and efficiency. Therefore, exploring optimization strategies targeting the complexity of self-attention remains a promising direction for future research;
- d. The performance gains observed in hierarchical structures for dense prediction tasks are particularly noteworthy when compared to recognition tasks, suggesting that developing a general hierarchical architecture for multiple vision tasks is a research direction worthy of further exploration.

In this section, we follow the sequence in which the input stream is forwarded within the ViT models to present the methods involved, offering the advantage of:

- a. By analyzing the performance bottlenecks of existing models, we can determine which component corresponds to lightweight measures. This section facilitates this process by providing a straightforward and accessible comparative framework.
- b. By emphasizing the relative independence of ViT components, this offers a perspective on adaptive and integrated optimization, which allows the selection of multiple components along the input stream to achieve lightweight design.

#### 4 Dynamic Network

Dynamic networks represent a critical evolution in the field of image recognition, dramatically improving precision by flexibly adapting their computational structures to the complexities of varying inputs. Unlike static models, which operate on fixed computational graphs and parameters, dynamic networks can tailor their internal configurations, thus achieving better computational efficiency and adaptability. This adaptability is particularly beneficial in image recognition tasks, where the complexity of images can vary widely. Furthermore, the dynamic network paradigm is one of the specific forms of online lightweighting strategies, where the computational load is reduced dynamically to maintain efficiency without compromising accuracy. In this section, we categorize dynamic networks based on the vision transformer into two primary topics to review the rapid advances: **Dynamic Resolution** in Section 4.1 and **Depth Adaptation** in Section 4.2.



Fig. 7. **Dynamic Resolution Exploration Timeline.** The figure collects and organizes the development history of dynamic resolution methods in ViT model lightweighting. It is mainly divided into two parts: token pruning and token merging.

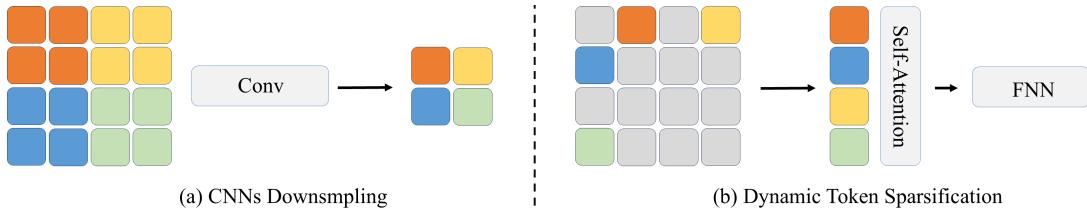
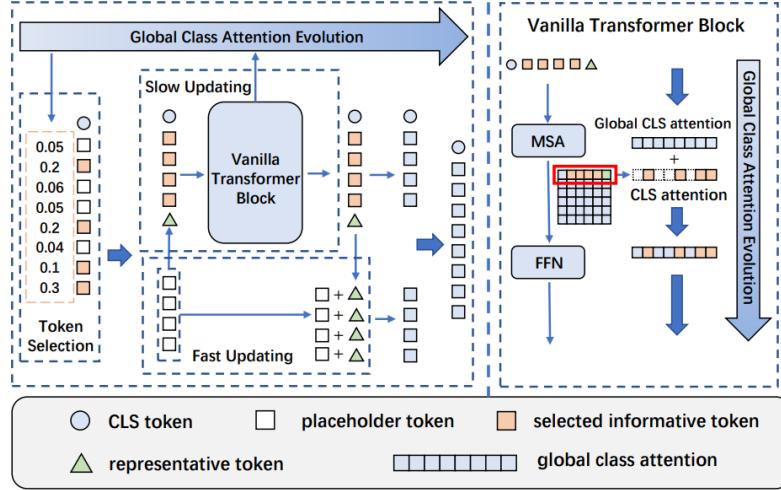


Fig. 8. **Space Redundancy Solution Comparison.** The CNN models typically employ a structural downsampling strategy, as illustrated in (a). The unstructured and data-dependent downsampling method in (b) more effectively capitalizes on the sparsity within the input data. Owing to the inherent characteristics of the self-attention operation, the unstructured token set is also accelerated through parallel computing.

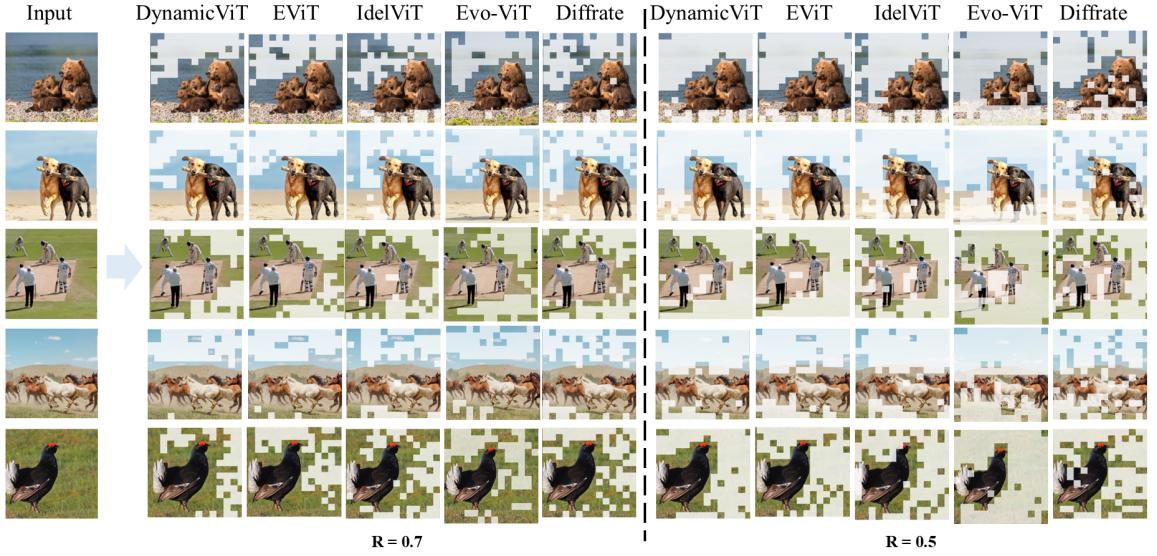
#### 4.1 Dynamic Resolution

As shown in Figure 8, unlike traditional convolutional neural networks, which often employ coarse sampling methods when faced with spatial redundancy or the need for downsampling, ViT models possess the unique token representation advantage, allowing for more customized and fine-grained downsampling strategies. Specifically, decisions are made about whether each token is pruned, kept, or merged based on its significance. This approach has proven particularly beneficial for image recognition tasks, where the ability to dynamically adjust the resolution and focus on significant regions of an image can greatly enhance the accuracy and efficiency of the model. From these applications, we summarize two main topics: **Dynamic Token Pruning** and **Dynamic Token Merging**. In Figure 7, we summarize the timeline of significant explorations on these two topics.

**4.1.1 Dynamic Token Pruning.** In relation to this topic, tokens with insufficient informational content are often directly pruned or sent through a different update path, whereas tokens recognized as having rich semantic information are passed to the deeper layers of the network.



**Fig. 9. Token Selection Example.** Evo-ViT [172] uses a token selector to pick informative tokens for slow updates and less-informative tokens for fast updates.



**Fig. 10. Token Pruning Visualization.** Visualize and compare different token pruning methods. Here,  $R$  represents the token keep rate, and the masked tokens are shown on a white background. Generally, the higher the attention to the object, the more effective the token pruning method is considered.

The Top-K pruning method [186] keeps only the  $K$  most significant tokens based on their attention weights. PS-ViT<sup>12</sup> [187] reduces tokens according to their measured importance to minimize computational costs while retaining performance. Patch thinning [89] employs a top-down strategy, with later layers guiding the token selection process in

<sup>12</sup><https://github.com/yuexy/PS-ViT>

Table 6. A quantitative comparative analysis is conducted on the classic methods involved in **Dynamic Token Pruning** (Section 4.1.1), using ImageNet-1K as the benchmark, with the resolution standardized to  $224 \times 224$ . The R in “/R” indicates the ratio of token pruning. The best-performing method is highlighted in bold, and the runner-up is underlined.

Model	Year	Params (M)↓	Throughput (img/s)↑	FLOPs (G)↓	Top-1 Acc. (%)↑
● Evo-ViT(DeiT-T) [172]	2021	5.9	4,027	0.8	72.0
● Evo-ViT(DeiT-S) [172]	2021	22.4	1,510	3.0	79.4
● Evo-ViT(DeiT-B) [172]	2021	87.3	462	11.6	81.3
● DynamicViT(DeiT-T)/0.7 [173]	2021	5.9	2,581	1.0	71.0
● DynamicViT(DeiT-T)/0.5 [173]	2021	5.9	2,590	0.8	67.4
● DynamicViT(DeiT-S)/0.9 [173]	2021	22.8	1,524	4.0	79.8
● DynamicViT(DeiT-S)/0.7 [173]	2021	22.8	2,062	2.9	79.3
● DynamicViT(DeiT-T)/0.7 [173]	2021	89.4	373	11.2	81.3
▲ EViT-Tiny [174]	2022	12.1	3,568	1.9	79.9
▲ EViT-Small [174]	2022	23.4	2,007	3.4	82.6
▲ EViT-Base [174]	2022	42.6	1,008	6.3	83.9
▲ EViT-Large [174]	2022	60.1	-	9.4	84.4
▲ SPViT(DeiT-T)/1.0 [175]	2022	5.7	1,372	1.3	72.2
▲ SPViT(DeiT-T)/0.7 [175]	2022	5.7	1,680	0.8	72.1
▲ SPViT(DeiT-S)/1.0 [175]	2022	22.0	718	4.6	79.9
▲ SPViT(DeiT-S)/0.7 [175]	2022	22.0	947	2.9	79.5
▲ SPViT(DeiT-B)/1.0 [175]	2022	86.6	246	17.7	82.2
▲ SPViT(DeiT-B)/0.7 [175]	2022	86.6	322	11.1	82.0

Model	Year	Params (M)↓	Throughput (img/s)↑	FLOPs (G)↓	Top-1 Acc. (%)↑
■ HeatViT-T0 [176]	2023	-	4,172	0.5	70.8
■ HeatViT-T2 [176]	2023	-	3,200	1.0	72.2
■ HeatViT-S0 [176]	2023	-	1,390	1.8	78.5
■ HeatViT-S1 [176]	2023	-	1,860	2.0	79.0
■ HeatViT-B0 [176]	2023	-	548	6.1	80.1
■ HeatViT-B1 [176]	2023	-	415	10.5	81.1
■ CF-ViT(DeiT-S)/1.0 [177]	2023	22.1	2,760	4.0	80.8
■ CF-ViT(DeiT-S)/0.75 [177]	2023	22.1	3,701	2.6	80.7
■ CF-ViT(DeiT-S)/0.5 [177]	2023	22.1	4,903	1.8	79.8
■ IdleViT(DeiT-S)/0.9 [178]	2023	22.1	2,662	4.0	79.9
■ IdleViT(DeiT-S)/0.7 [178]	2023	22.1	3,361	3.1	79.6
■ IdleViT(DeiT-S)/0.5 [178]	2023	22.1	4,071	2.4	79.0
▲ ATS(DeiT-S) [179]	2022	22.0	1,403	2.9	79.7
▲ ATS(CvT-13) [179]	2022	22.0	1,080	3.2	81.4
■ TPC-ViT(DeiT-T) [180]	2024	5.7	4,500	0.6	73.0
■ TPC-ViT(DeiT-S) [180]	2024	22.0	2,303	2.8	80.8
■ TPC-ViT(DeiT-B) [180]	2024	86.6	490	10.7	81.8

Table 7. A quantitative comparative analysis is conducted on the classic methods involved in **Dynamic Token Merging** (Section 4.1.2), using ImageNet-1K as the benchmark, with the resolution standardized to  $224 \times 224$ . The R in “/R” shows the token merging rate. If  $R < 1$ , the model keeps R% of the original tokens. If  $R > 1$ , the model is tested after merging tokens into R blocks. The best-performing method is highlighted in bold, and the runner-up is underlined.

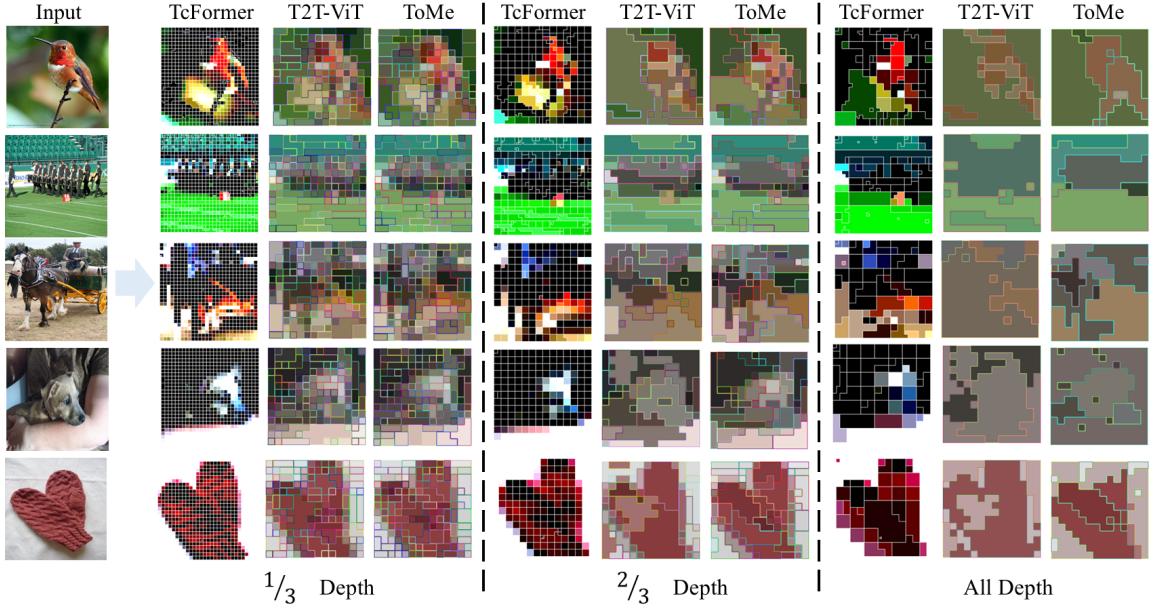
Model	Year	Params (M)↓	Throughput (img/s)↑	FLOPs (G)↓	Top-1 Acc. (%)↑
● T2T-ViT(ViT-14) [92]	2021	21.5	1,944	4.8	81.5
● T2T-ViT(ViT-19) [92]	2021	39.2	1,660	8.5	81.9
● T2T-ViT(ViT-24) [92]	2021	64.1	1,497	13.8	82.3
■ MHSA(DeiT-T) [181]	2023	5.8	-	0.8	72.7
■ MHSA(DeiT-S) [181]	2023	22.2	-	3.1	79.7
■ MHSA(LV-ViT-S) [181]	2023	26.2	-	3.9	82.8
■ MHSA(DeiT-B) [181]	2023	86.6	-	11.5	81.5
■ SuperViT(LV-ViT-T)/1.0 [182]	2023	5.5	3,178	2.9	79.9
■ SuperViT(DeiT-S)/0.5 [182]	2023	5.7	4,767	2.3	80.0
■ SuperViT(DeiT-T)/1.0 [182]	2023	5.7	5,013	1.3	73.1
■ SuperViT(DeiT-S)/0.7 [182]	2023	22.3	3,654	3.0	80.5
■ SuperViT(DeiT-S)/1.0 [182]	2023	22.3	2,461	4.6	80.6
■ SuperViT(LV-ViT-S)/0.7 [182]	2023	27.8	2,684	4.3	83.2
■ SuperViT(LV-ViT-S)/1.0 [182]	2023	27.8	1,748	6.6	83.5
■ ToMe(ViT-B)/8 [183]	2023	86.0	413	13.1	82.9

Model	Year	Params (M)↓	Throughput (img/s)↑	FLOPs (G)↓	Top-1 Acc. (%)↑
■ ToMe(ViT-B)/12 [183]	2023	86.0	489	10.9	81.8
■ ToMe(ViT-B)/16 [183]	2023	86.0	607	8.7	78.9
■ ToMe(ViT-B)/20 [183]	2023	86.0	736	7.1	67.5
■ ToMe(ViT-L)/8 [183]	2023	307.0	182	61.6	<b>84.2</b>
■ ToFu AVG(ViT-B)/8 [184]	2023	86.6	417	13.1	83.2
■ ToFu AVG(ViT-B)/12 [184]	2023	86.6	484	10.9	82.4
■ ToFu AVG(ViT-B)/16 [184]	2023	86.6	615	8.7	80.4
■ ToFu AVG(ViT-B)/20 [184]	2023	86.6	748	7.1	72.2
▲ TCFormerV1-Light [185]	2024	14.2	185	3.8	79.6
▲ TCFormerV2-Tiny [185]	2024	14.2	393	2.5	79.5
▲ TCFormerV1 [185]	2024	25.6	120	5.9	82.4
▲ TCFormerV2-Small [185]	2024	25.6	273	4.5	82.4
▲ TCFormerV1-Large [185]	2024	62.8	58	12.2	83.6
▲ TCFormerV2-Base [185]	2024	62.8	103	10.8	<b>83.8</b>

earlier layers. EViT<sup>13</sup> [174] fuses less important tokens into a single token for further computation, preserving the essence while reducing the number. Evo-ViT [172] is shown in Figure 9, regulates the flow of gradient propagation, updating critical tokens slowly and less important ones quickly before merging. IdleViT [178] implements a token idling strategy, processing only a subset of tokens in each layer and passing the rest to the next layer unchanged.

Recent studies emphasize the importance of token selection mechanisms adapted to evolving ViT architectures. SPViT [175] employs a dynamic multi-head token selector to filter and integrate less informative tokens. The token compensator (ToCom) [188] proposes a model arithmetic framework to decouple compression degrees between the

<sup>13</sup><https://github.com/youweiliang/evit>



**Fig. 11. Token Merge Visualization.** Visualize the comparison of token merging methods, where “1/3”, “2/3”, and “All” indicates the stage within the ViT stack structure where the token merging operation is executed. Typically, the shallower layers focus on merging detailed texture information, while the deeper layers attend to the merging of more abstract semantic information.

training and inference stages, addressing performance drops caused by mismatched compression degrees, and achieving universal performance improvements without further training. HeatViT [176] introduces a delay-aware, multi-stage training strategy to optimize token selector performance. TSNet [189] dynamically selects informative tokens from video samples using a lightweight scoring module, while DynamicViT [93] incorporates a learnable module within each transformer block for layer-by-layer token pruning. TPC [180] uses suspension and reuse probabilities in token selection to preserve computational accuracy. Lastly, the Adaptive Token Sampler (ATS) [179] offers a differentiable parameter-free method for adaptive downsampling, minimizing information loss. In Figure 10, we present a visual comparison of token pruning strategy methods, generally indicating that the degree of focus on objects demonstrates the ability of the method to identify regions crucial to decision-making.

**4.1.2 Dynamic Token Merging.** Unlike pruning methods that discard tokens directly, token merging through strategic token merging, this paradigm can combine less informative tokens with more critical ones, reduce the number of tokens processed, thus contributing to the broader goal of model simplification.

ToMe [183]<sup>14</sup> introduces a fine-tuned matching algorithm to merge similar tokens across layers. PatchMerger [190] allows the network to handle varying token counts by decoupling the number of output tokens from the input, increasing flexibility. [191] presents the token pruning and squeezing (TPS) module, which uses a token merging strategy based on nearest neighbor matching and similarity. MHSA [181] proposes merging less critical tokens before pruning to maintain crucial information and prevent performance drops. [192] devises a meta-token mechanism to generate an attention map to merge similar tokens. T2T-ViT [193] rearranges the input token sequence to focus on important local

<sup>14</sup><https://github.com/facebookresearch/ToMe>

structures, facilitating effective token merging. TCFormer [185] uses progressive clustering to flexibly merge tokens from various locations, capturing detailed information. Lastly, Token Fusion [184] dynamically combines pruning and merging based on token computation. Figure 11 presents a visual comparison of token merging strategies, illustrating how merging is performed at varying depths within the ViT architecture. It is especially noticeable that the initial layers primarily merge fine-grained texture details, whereas the subsequent layers are more focused on integrating higher-level semantic information.

**4.1.3 Summary: Dynamic Resolution.** Eliminating spatial redundancy enables ViT to offer lightweight online design. By evaluating token importance, it retains, prunes, or merges tokens, ideal for real-time image recognition on high-resolution images or large-scale datasets. Relevant experimental analyzes are provided in Table 6 and Table 7, from which it can be observed that: **(i)**. The dynamic resolution strategy can still achieve a lightweight extreme. For example, Evo-ViT [172] has only 5.9M parameters and has a high throughput. **(ii)**. Most methods can significantly improve throughput and reduce FLOPs while slightly decreasing accuracy. For example, DynamicViT [93] increases throughput by 36.6% and reduces FLOPs by 27.5% compared to the original DeiT-S model, with only the 0.5% accuracy drop. **(iii)**. Larger models (such as ViT-B, ViT-L) can typically tolerate more aggressive token reduction while maintaining high accuracy. For example, ToMe [183] substantially reduces computational cost while maintaining a high accuracy of 84.2%. **(iv)**. Different methods have their own advantages in balancing performance and efficiency. For example, the EViT [194] series excels in maintaining high accuracy, while the SuperViT [182] series has more advantages in improving throughput.

## 4.2 Depth Adaptation

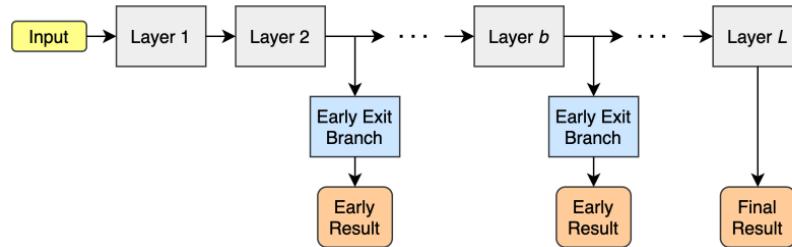


Fig. 12. **Early Exit Paradigm.** Attaching additional internal classifiers at various stacked layers of the ViT model halts computation upon satisfying predefined exit conditions, thereby facilitating the early exit of simpler samples without compromising image recognition accuracy, which in turn conserves computational resources.

Dynamic multi-exit networks provide a flexible online lightweighting technique by enabling early computation halts. AViT [195] adapts ViT to dynamically adjust the depth and computation of the layer based on the specific task and the complexity of the input data. [196] enhances this paradigm using the knowledge distillation training procedure, allowing early exits to benefit from the abstract semantic knowledge of deeper layers. As shown in Figure 12,[197] introduces a novel early exit architecture that uses fine-tuning to significantly improve exit accuracy. [198] integrates multi-exit architecture into ViT, proposing seven distinct early exit designs. To address performance degradation in early exit methods, LGViT [199] takes advantage of heterogeneous exit headers (local perceptual and global aggregation) to balance efficiency and precision. DYN-ADAPTER<sup>15</sup> [200] dynamically adjusts the model capacity based on disentangled

<sup>15</sup><https://github.com/LMD0311/DAPT>

representations. CF-ViT<sup>16</sup> [177] divides the processing pipeline into coarse and fine steps. AdaViT<sup>17</sup> [201] adapts the depth adjustment of ViT models based on the complexity of the task.

To further enhance the efficiency and adaptability of the ViT models, several dynamic training strategies are explored. [202] presents a faster depth-adaptive transformer, taking advantage of dynamic adjustment of the model layers to optimize computational resources. [203] allows different layers to exit at different points, thus reducing unnecessary computations. [204] proposes a confident adaptive language model that dynamically halts computations based on confidence measures, which can be paralleled in vision tasks to enhance efficiency. Additionally, Skipdecode [205] presents an autoregressive skip decoding mechanism that can potentially be adapted for ViT models to optimize inference efficiency by skipping certain decoding steps while maintaining performance.

**4.2.1 Summary: Depth Adaptation.** In traditional fixed-depth networks, both simple and complex inputs undergo the same computational process, which to some extent wastes resources. This topic introduces a dynamic multi-exit paradigm that allows the model to flexibly adjust its computational depth based on input complexity, which aligns with the principles of online lightweighting strategies. The multi-exit structure can be regarded as a form of “implicit ensemble”, where models of different depths share parameters but provide diverse decisions. Moreover, through knowledge distillation techniques, the early layers can also acquire deep semantic information. As ImageNet is a large-scale and inconsistent difficulty dataset, dynamic depth can offer better adaptability, improving overall recognition accuracy. For simple images, results can be obtained at shallower layers, reducing unnecessary computations.

### 4.3 Dynamic Networks Discussion

Based on the complexity of the input and the inherent token representation advantages of ViT, a wide variety of dynamic networks can be constructed. Compared to the dynamic network paradigm of the CNN model, ViT has the additional advantage of handling input sequences of variable length. This means ViT does not need to worry about resizing to a rectangular space after eliminating redundancy, thus offering greater flexibility. This flexibility is particularly beneficial in lightweight online strategies, where token pruning and merging are employed to reduce computational load while simultaneously enhancing recognition accuracy. We report the relevant experimental results in Table 6 and Table 7, and summarize the limitations of the existing methods as follows:

- a. Lack of guidance for the pruning or merging layers, with the current widespread practice of manually setting them to be executed at the 3<sup>th</sup>, 6<sup>th</sup>, and 9<sup>th</sup> layers[93, 194].
- b. Adaptive pruning or merging rates, although learnable pruning rate methods such as DiffRate [206] are now available, require additional complex optimization techniques and training overhead.
- c. Token pruning or token merging needs to be adapted to subsequent model pruning or quantization, making an end-to-end training strategy necessary.
- d. Although ViT can handle token sequences of variable lengths, token pruning or token merging disrupts position information, necessitating methods for re-modeling position information.

On the other hand, for depth adaptation, the basic paradigm of the early exit strategy in CNN is generally followed, but the following limitations still exist:

- a. Internal classifiers serve primarily as projections to category space, failing to fully leverage the advantage of token representations.

<sup>16</sup><https://github.com/ChenMnZ/CF-ViT>

<sup>17</sup><https://github.com/MengLcool/AdaViT>

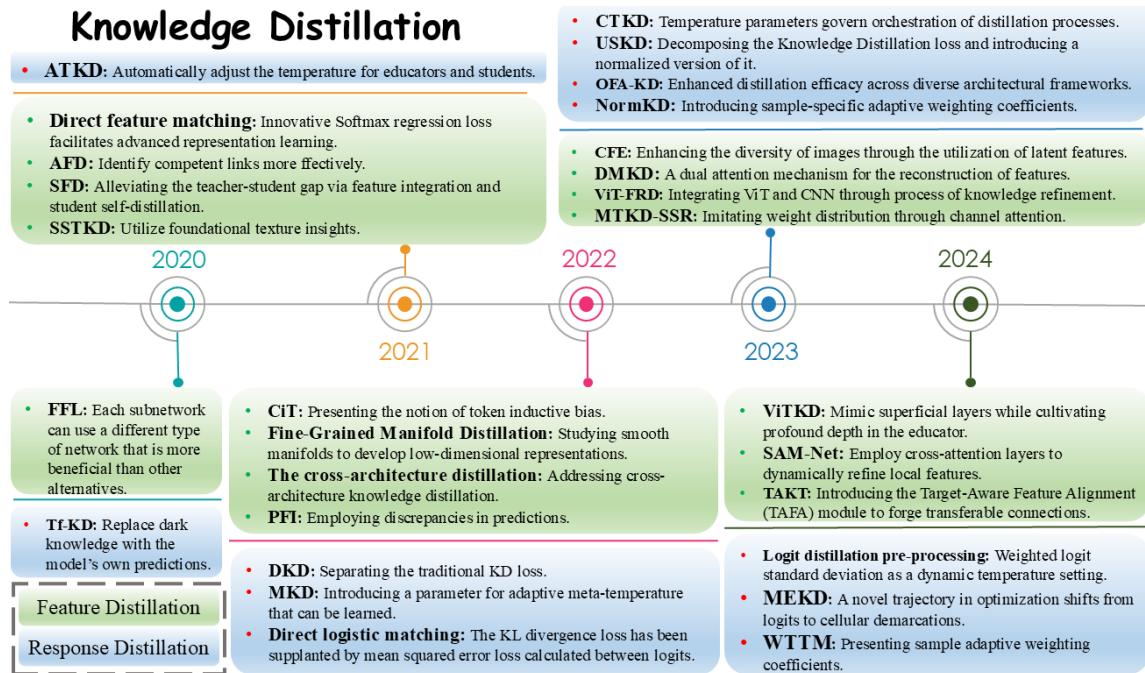


Fig. 13. **Knowledge Distillation Exploration Timeline.** The figure collects and organizes the development of knowledge distillation methods in ViT model lightweighting. It includes two parts: feature knowledge distillation and response knowledge distillation.

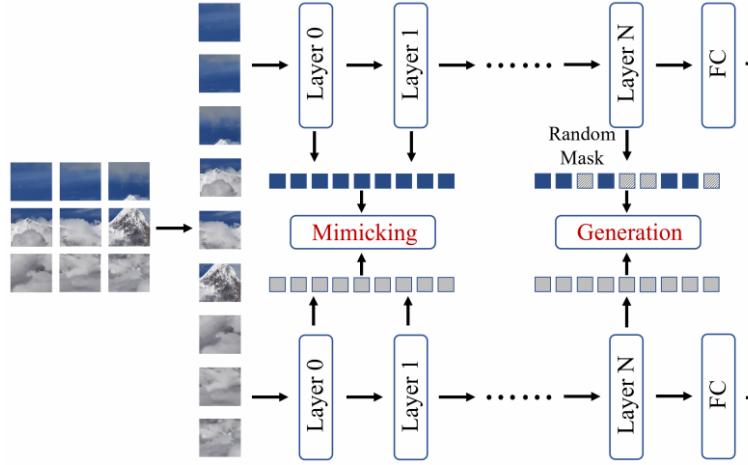
- Lack of effective guidance for the placement of internal classifiers.
- The potential to skip individual components within transformer blocks, such as multi-head self-attention modules or feed-forward networks, rather than entire blocks.

## 5 Knowledge Distillation

Knowledge distillation has emerged as a powerful approach to compress complex ViT models into lightweight and efficient variants suitable for image recognition tasks. By extracting and transferring knowledge from a large, pre-trained teacher model to a smaller student model, knowledge distillation enables the creation of compact ViT architectures that maintain high accuracy while significantly reducing storage requirements and inference time. [207, 208] propose the concept of knowledge distillation, mainly focused on extracting class probability distributions through softening of labels. In the context of lightweight online ViT strategies, knowledge distillation can be broadly categorized into two main topics: **Feature Knowledge Distillation** in Section 5.1 and **Response Knowledge Distillation** in Section 5.2. In Figure 13, we summarize the timeline of significant explorations on these two topics. It should be noted that the relation distillation, present in CNN models, remains applicable, while the relation distillation among tokens in ViT is more adapted.

### 5.1 Feature Knowledge Distillation

Feature-based knowledge distillation (FKD) focuses on transferring rich intermediate-level information from a teacher model to a student model, capitalizing on the detailed representations within the internal layers of teacher networks. FKD



**Fig. 14. Feature Knowledge Distillation Example.** ViTKD [215] observes that regardless of model capacity, the attention gap in shallow layers is relatively small, advocating for a direct alignment strategy. In contrast, the attention distribution in deeper layers is more diverse, thus necessitating a generative alignment approach.

employs the multi-faceted strategy: **Attention Feature Distillation**, captures essential attention patterns, **Structured Feature Distillation**, compresses structural representations, and **Direct Feature Mimicking**, ensures that important features are transferred directly.

FKD emphasizes comprehensive guidance through intermediate feature information. The typical loss function for feature-based distillation is defined as:

$$\mathcal{L}_{\text{FKD}}(F^S, F^T) = \mathcal{L}_D(\phi^S(F^S), \phi^T(F^T)), \quad (3)$$

where  $F^S$  and  $F^T$  denote the intermediate feature mappings from the student and teacher, respectively. The functions  $\phi^S$  and  $\phi^T$  transform these features to include enhanced information such as attention mechanisms, activation boundaries, neuron selectivity, and probability distributions. The loss function  $\mathcal{L}_D$  measures the similarity between features, using metrics such as mean square error and Kullback-Leibler divergence.

**5.1.1 Attention Feature Distillation.** Attention feature distillation transfers valuable information from teacher model attention maps. AttnDistill [209] aligns class labels using the Projector Alignment (PA) module and approximates teacher network attention maps with the Attention Guidance (AG) module. PFI [210] guides feature mimicry by emphasizing high-variance areas based on prediction variance. ViT-FRD [211] refines knowledge transfer by integrating ViT and CNN using the Collaborative Reformer based on Clustering (CCReformer), the Linear Transformer (Linformer) as the knowledge distillation modules. CFD [212] employs a dataset-independent classifier to distill the distribution of refined features to the student. AFD<sup>18</sup> [213] takes advantage of attention-based meta-networks to control the strength of distillation by identifying similarities. SAM-Net<sup>19</sup> [214] dynamically adapts local features based on environmental and spatial correlations by exploiting cross-attention. Attention Probe-Based Distillation [209] selects significant real-world data and uses a probe knowledge distillation algorithm to train lightweight student models.

<sup>18</sup><https://github.com/holger24/AFD>

<sup>19</sup><https://github.com/benjaminkeleenyi/SAM-Net>

**5.1.2 Structured Feature Distillation.** Structured feature distillation ensures effective knowledge transfer while maintaining structural integrity, which is crucial for lightweight vision transformers. Preserves and transfers structural information, helping student networks capture complex patterns and relationships, thus improving performance and generalization.

KDFAS [216] addresses the issue of embedding dimension gap by employing a covariance matrix during feature-level knowledge transfer, while DMKD [217] employs a dual-attention mechanism to guide spatial and channel information to its respective masking branches and integrates reconstructed features. PromptKD [218] transfers category embedding from teacher to student through shared distillation. CrossKD [219] makes the student generate cross head predictions like the teacher. MTKD-SSR<sup>20</sup> [220] introduces the Staged Channel Distillation (SCD) mechanism, which takes advantage of channel attention to allow the student network to mimic the weight distribution within the channel feature map. PSD<sup>21</sup> [221] incorporates multiple self-distillations with a shared embedded network between the teacher and the student during each process, and TAKT [222] proposes the Target-Aware Feature Alignment (TAFA) module to establish transferable feature connections between the source and target domains. The cross-architecture knowledge distillation approach [223] aligns intermediate features using projectors for partial cross-attention and grouped linear transformations, while Fine-Grained Manifold Distillation [224] creates low-dimensional features by learning smooth manifolds embedded within the original feature space. DearKD [225] is a two-stage framework that initially refines the inductive bias from the early intermediate layers of the CNN, subsequently allowing the transformer to be fully leveraged through unrefined training.

**5.1.3 Direct Feature Mimicking.** Direct Feature Mimicking alleviates the need for complex preprocessing of alignment targets, helping to preserve the native quality and inherent patterns of the information [226]. As shown in Figure 14, ViTKD<sup>22</sup> [215] employs “linear layers” and “correlation matrices” for shallow tasks, a convolutional projector for deeper generative tasks. MMKD [227] uses a meta-weighting network to integrate logic and intermediate features from various teacher models to guide the student network. [228] proposes a one-to-one spatial matching method that employs a target-aware transformer<sup>23</sup> (TaT) to align the feature component.

**5.1.4 Summary: Feature Knowledge Distillation.** To improve the performance of small capacity ViT models and facilitate their deployment in constrained environments, directly training compact ViT models often fails to achieve satisfactory performance, and simple knowledge distillation methods are insufficient to fully transfer the rich information embedded within the teacher model.

Therefore, extracting the abundant latent patterns from the teacher network presents a promising direction. Attention Feature Distillation enhances the ability of lightweight ViT models to identify critical information, while Structured Feature Distillation maintains the integrity of knowledge structures during the distillation process. Direct Feature Mimicking alleviates the burden of complex preprocessing. This multifaceted feature knowledge distillation strategy effectively transfers the rich intermediate-layer information. As illustrated in Figure 15, TinyMIM<sup>24</sup> [229] takes advantage of the unique representational advantages inherent in ViT architectures and extracts relation knowledge among tokens as a complementary feature knowledge transfer, which differs from the conventional CNN approach of extracting relation knowledge among images within the input batch.

<sup>20</sup><https://github.com/ChaofWang/Awesome-Super-Resolution>

<sup>21</sup><https://github.com/webtoon/psd>

<sup>22</sup>[https://github.com/yzd-v/cls\\_KD](https://github.com/yzd-v/cls_KD)

<sup>23</sup><https://github.com/Kevoen/TATTrack>

<sup>24</sup><https://github.com/OliverRensu/TinyMIM>

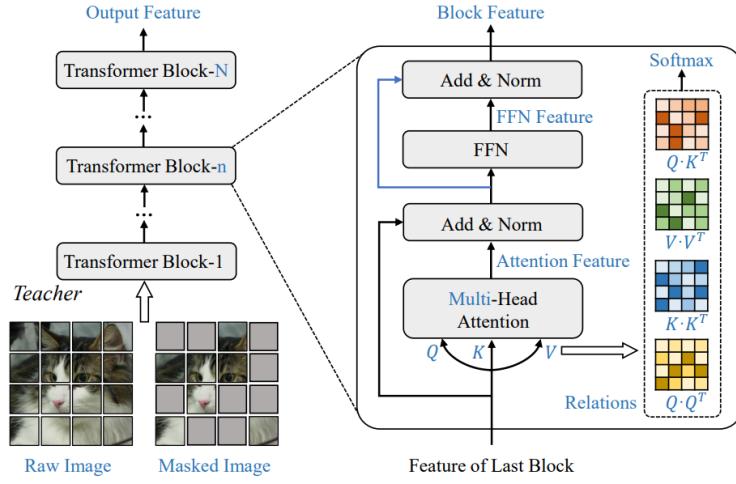


Fig. 15. **Token Relation Transfer.** TinyMIM [229] extracts the relation knowledge among tokens, including  $Q \cdot K^T$ ,  $V \cdot V^T$ ,  $K \cdot K^T$ , and  $Q \cdot Q^T$ , which differs from the conventional CNN of capturing the relation knowledge among images within the batch.

## 5.2 Response Knowledge Distillation

Feature-based approaches recently come to the fore due to its superior performance in numerous applications. Consequently, response knowledge distillation (RKD) is relatively neglected. RKD, which facilitates direct learning from the output layer of the teacher model by the student model, presents distinct advantages, including rapid learning and straightforward implementation, and the potential remains largely unexplored. This outcome-driven approach makes it widely applicable across various tasks, given the recognition task, the soft probability distribution  $p$  is defined as:

$$p(z_i; T) = \frac{\exp(z_i/T)}{\sum_{j=1}^N \exp(z_j/T)}, \quad (4)$$

where  $z_i$  is the logit for class  $i$ ,  $N$  is the number of classes, and  $T$  is a temperature parameter controlling the distribution smoothness. RKD aims to align the teacher probability distribution,  $p(z^T; T)$ , with the student,  $p(z^S; T)$ , by minimizing a distance function  $\mathcal{L}_{dis}$ :

$$\mathcal{L}_{RKD} = \mathcal{L}_{dis} \left( p(z^S; T), p(z^T; T) \right), y \quad (5)$$

where  $\mathcal{L}_{dis}$  can be instantiated with various distance metrics, such as Kullback-Leibler divergence, mean squared error, or Pearson correlation. For example, the classic KD loss is formulated as:

$$\mathcal{L}_{RKD} = \sum_{i=1}^N p(z_i^T; T) \log \frac{p(z_i^T; T)}{p(z_i^S; T)}, \quad (6)$$

where  $z_i^T$  and  $z_i^S$  represent the teacher and student logits for class  $i$ , respectively.

Improving response-based knowledge distillation can involve adjusting distillation **Temperature Variants** to transfer information-rich knowledge, optimizing **Objective Function Design** to diversify learning objectives, and **Empowering Soft Labels** to prevent overfitting.

**5.2.1 Temperature Variants.** Temperature variation by modulating the softness of the teacher model logit output, temperature scaling provides smoother and more informative probability distributions to the student model. However,

traditional fixed-temperature methods often fail to exploit the diverse properties of logit distributions across samples, leading to suboptimal knowledge transfer [230].

To overcome these limitations, there are several advanced techniques. ATKD [231] dynamically adjusts the temperatures for both teacher and student outputs, balancing clarity and mitigating performance degradation. MKD [232] introduces meta-learned temperature parameters that adapt throughout the training process based on the learning objective gradient. [233] introduces a multi-level logit distillation framework, which transforms them into multiple outputs with varying temperatures, and performs multi-level alignment matching. NormKD [234] customizes the temperatures for each sample to achieve a uniform distribution by analyzing logit output. SKD [235] integrates a temperature-scaled LogSoftmax function with a learning simplifier to improve distillation loss during joint training. [236] propose using a weighted logit standard deviation as the adaptive temperature, combined with Z-score preprocessing to normalize logits before applying Softmax.

**5.2.2 Objective Function Design.** Creative formulations of objective functions enhance the distillation process, allowing a wider variety of training for compact student models.

Decoupled Knowledge Distillation (DKD) [237] splits KD loss into Target Class (TCKD) and Non-Target Class (NCKD) components, with NCKD weighting independent of teacher confidence. Unified Self-Knowledge Distillation (USKD) [238] further refines the KD loss into target and non-target losses using cross-entropy, normalizes non-target logits to formulate the non-target knowledge distillation (NKD) loss, and generates customized soft labels to potentially improve knowledge transfer. PTLoss [239] represents the original KL-based distillation loss function through a maclaurin series and implicitly transforms the original teacher into a proxy teacher by perturbing the first-order term in this series. NTCE [240] introduces a Magnitude Kullback Leibler (MKL) divergence to better represent non-target categories, combined with diversity-based data augmentation (DDA) to enrich non-target category diversity. CSKD [71] terminates the pooling operation following the final feature map, producing dense predictions for each corresponding feature location. From these results, hard labels are derived to serve as target labels, utilizing cross-entropy as the loss function for spatial knowledge transfer.

**5.2.3 Empowering Soft Labels.** Reducing the influence of the teacher model in knowledge distillation seeks to create vision transformer models that are more lightweight and autonomous. This approach moves away from a strong reliance on pre-trained teachers, allowing students to learn on their own.

The Tf-KD framework [241] replaces dark knowledge with the model own predictions, using self generated soft targets for regularization, and achieves smoothing effects similar to label smoothing regularization (LSR). MTKDSSR [220] incorporates staged channel distillation, staged response distillation, and cross-stage distillation, using student self-reflection through self-distillation of response-based knowledge. The OFA-KD framework [242] eliminates architecture-specific information and employs exit branches in the student model trained with teacher model logits. MEKD [243] transitions from logits to cell boundaries, innovating with de-privatization and distillation for lightweight model transformations. WTTM [244] enhances the training of the student model using sample-adaptive weighting to emulate the power-transformed probability distributions of the teacher. Tiny-ViT [245] focuses on reducing memory and computational overhead through logit sparsification and efficient knowledge transfer during pre-training.

**5.2.4 Summary:Response Knowledge Distillation.** Response knowledge distillation enhances efficiency, prevents overfitting, and facilitates cross-architecture knowledge sharing due to its flexible implementation [246]. We cover three topics: Temperature Variants exploit logit distribution characteristics better than fixed temperature methods; Objective

Function Design diversifies learning objectives for more comprehensive knowledge transfer; Empowering Soft Labels lessens dependence on pre-trained teacher models, enhancing student models independent learning. Furthermore, *(i)*. Response knowledge distillation provides soft guidance, offering richer information than hard labels. *(ii)*. Given the prevalent long-tail issue in image recognition tasks, leveraging the teacher model performance on rare categories to improve the student model’s learning can be achieved by adjusting the response distillation loss weight to focus on rare categories. *(iii)*. Provides a lightweight response-based multi-scale knowledge transfer pathway [247], distinctly different from the burdensome multi-scale feature knowledge distillation.

### 5.3 Knowledge Distillation Discussion

Knowledge distillation bridges CNN and ViT models for cross-structural knowledge transfer. We focus on KD strategies that take advantage of both, especially ViT. For example, AttnDistill [209] uses the importance distribution between the [CLS] token and the image tokens to achieve a more efficient alignment of the target. ViT models effectively extract relevant knowledge for tasks in response knowledge distillation. TinyMIM [229] adapts the popular relation knowledge transfer in CNN to ViT using Q-K, V-V, K-K, and Q-Q relationships among tokens. We summarize the limitations of existing methods as follows:

- a. Exploiting the unique structural advantages of ViT for knowledge distillation, including head-level distillation, where existing methods still involve numerous artificial settings or task-coupling characteristics.
- b. The fast distillation direction, focusing on response knowledge distillation, can be extended to feature knowledge distillation and even combined with self-supervised optimization objectives.
- c. Addressing privacy and security concerns during the knowledge distillation process as well as enhancing the robustness of the student network in abnormal scenarios post-distillation.

## 6 Future Research Directions

Vision transformer models excel in computer vision, but their high parameter and computational needs hinder their practical use. Future lightweight ViT models will emphasize efficiency, intelligence, and security.

**Representation Advantage Expansion.** Breaking free of the limitations of static model structures, the focus shifts from the previous global perspective to a more fine-grained local perspective, while integrating the unique representational advantages of ViT, such as deep adaptability and response knowledge distillation, which remain directions worthy of profound exploration in their ingenious integration with ViT.

**End-to-End Training System.** Offline lightweight ViT techniques, such as post-training quantization, remain relatively independent of online lightweight techniques. Future exploration can focus on designing end-to-end training systems, where errors arising from token pruning and merging can also be perceived and controlled.

**Automatic Configuration.** Existing online lightweight strategies for ViT still involve a significant number of scenarios that require manual intervention. We advocate for the automation of these processes, focusing primarily on two aspects:

- a. Analysis of performance bottlenecks and selection of the corresponding combination strategies can be automated by structuring the pipeline.
- b. Within specific strategies, manually set parameters, such as the location of the layer and the corresponding rate, remain for token pruning and merging.

**Privacy-Sensitive Lightweight.** Traditional knowledge distillation requires access to the original training dataset, which may not be feasible in privacy-sensitive application scenarios. Data-Free Knowledge Distillation (DFKD) methods Manuscript submitted to ACM

generate synthetic data using only the teacher model, without accessing the original data, thus protecting data privacy. However, DFKD has not yet been explored in the ViT domain. We offer several valuable research directions:

- a. Using the diffusion vision transformer [248, 249] instead of the CNN generator to transfer knowledge on synthetic datasets that better match real data distributions.
- b. The current DFKD setup [250] still reveals the original dataset input resolution to the user, and using the flexible ViT input form can enhance the benefits of the data-free paradigm.

**Inference Security.** Well-trained teacher networks possess greater robustness against anomalous inputs, but the lightweight optimization process of the ViT model does not explicitly transfer this knowledge. This makes small ViT models more susceptible to adversarial attacks. Consequently, the extraction and transfer of robust, interference-resistant knowledge is a crucial consideration for future practical deployments.

## 7 Conclusion

In this paper, we conduct a comprehensive investigation into lightweight techniques for vision transformer in image recognition, with particular emphasis on online lightweight strategies. Our exploration is divided into three key topics: vision transformer **Efficient Component Design**, **Dynamic Network**, and **Knowledge Distillation**. We analyze these lightweight approaches both quantitatively and qualitatively in terms of the trade-offs between precision and efficiency, under the consistent view on ImageNet-1K. For each topic, we present the methods in the form of input flow forward the network, allowing flexible component combinations and bottleneck identification. In addition, we discuss the limitations of related explorations within each topic and propose potential solutions. However, numerous unknown challenges remain in the face of complex real-world scenarios, thus advocate for maintaining the privacy and security of the lightweighting process to ensure the broad acceptance and usage of ViT. In summary, we hope that this survey will serve as a valuable resource for researchers, providing insight and guidance for the development of lightweight vision systems.

## References

- [1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]. <https://arxiv.org/abs/2010.11929>
- [2] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in Vision: A Survey. *ACM Comput. Surv.* 54, 10s, Article 200 (Sept. 2022), 41 pages. <https://doi.org/10.1145/3505244>
- [3] Shibiao Xu, Shuchen Zheng, Wenhao Xu, Rongtao Xu, Changwei Wang, Jiguang Zhang, Xiaoqiang Teng, Ao Li, and Li Guo. 2024. HCF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection. *ArXiv* abs/2403.10778 (2024). <https://api.semanticscholar.org/CorpusID:268512941>
- [4] Rongtao Xu, Changwei Wang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. 2023. Wave-Like Class Activation Map With Representation Fusion for Weakly-Supervised Semantic Segmentation. *IEEE Transactions on Multimedia* (2023).
- [5] Rongtao Xu, Changwei Wang, Jiaxi Sun, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. 2023. Self Correspondence Distillation For End-to-End Weakly-Supervised Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [6] Changwei Wang, Rongtao Xu, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. 2023. CNDesc: Cross Normalization for Local Descriptors Learning. *IEEE Transactions on Multimedia* 25 (2023), 3989–4001. <https://api.semanticscholar.org/CorpusID:248330052>
- [7] Changwei Wang, Rongtao Xu, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. 2023. Treating Pseudo-labels Generation as Image Matting for Weakly Supervised Semantic Segmentation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 755–765. <https://api.semanticscholar.org/CorpusID:267027139>
- [8] Changwei Wang, Shunpeng Chen, Yukun Song, Rongtao Xu, Zherui Zhang, Jiguang Zhang, Haoran Yang, Yu Zhang, Kexue Fu, Shide Du, Zhiwei Xu, Longxiang Gao, Li Guo, and Shibiao Xu. 2025. Focus on Local: Finding Reliable Discriminative Regions for Visual Place Recognition. In *AAAI Conference on Artificial Intelligence*. <https://api.semanticscholar.org/CorpusID:277764714>

- [9] Rongtao Xu, Changwei Wang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. 2025. RML: Efficient Representation Mutual Learning Framework for End-to-End Weakly-Supervised Semantic Segmentation. *IEEE Transactions on Instrumentation and Measurement* (2025). <https://api.semanticscholar.org/CorpusID:276948471>
- [10] Changwei Wang, Rongtao Xu, Shibiao Xu, Weiliang Meng, Ruisheng Wang, and Xiaopeng Zhang. 2024. Exploring Intrinsic Discrimination and Consistency for Weakly Supervised Object Localization. *IEEE Transactions on Image Processing* 33 (2024), 1045–1058. <https://api.semanticscholar.org/CorpusID:267256165>
- [11] Rongtao Xu, Changwei Wang, Shibiao Xu, Weiliang Meng, Yuyang Zhang, Bin Fan, and Xiaopeng Zhang. 2024. DomainFeat: Learning Local Features With Domain Adaptation. *IEEE Transactions on Circuits and Systems for Video Technology* 34 (2024), 46–59. <https://api.semanticscholar.org/CorpusID:259553858>
- [12] Rongtao Xu, Jiguang Zhang, Jiaxi Sun, Changwei Wang, Yifan Wu, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. 2024. MRFTrans: Multimodal Representation Fusion Transformer for monocular 3D semantic scene completion. *Information Fusion* (2024), 102493.
- [13] Zhen Jia, Zhang Zhang, Liang Wang, and Tieniu Tan. 2024. Human Image Generation: A Comprehensive Survey. *ACM Comput. Surv.* 56, 11, Article 279 (June 2024), 39 pages. <https://doi.org/10.1145/3665869>
- [14] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (*UIST ’23*). Association for Computing Machinery, New York, NY, USA, Article 38, 17 pages. <https://doi.org/10.1145/3586183.3606735>
- [15] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. 2024. DiffiT: Diffusion Vision Transformers for Image Generation. [arXiv:2312.02139 \[cs.CV\]](https://arxiv.org/abs/2312.02139) <https://arxiv.org/abs/2312.02139>
- [16] Shiv Ram Dubey and Satish Kumar Singh. 2024. Transformer-based generative adversarial networks in computer vision: A comprehensive survey. *IEEE Transactions on Artificial Intelligence* (2024).
- [17] Dmitrii Torbunov, Yi Huang, Haiwang Yu, Jin Huang, Shinjae Yoo, Meifeng Lin, Brett Viren, and Yihui Ren. 2023. Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 702–712.
- [18] Mohamed Amine Marnissi and Abir Fathallah. 2023. GAN-based vision Transformer for high-quality thermal image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 817–825.
- [19] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. 2023. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23164–23173.
- [20] Weilun Feng, Chuanguang Yang, Zhulin An, Libo Huang, Boyu Diao, Fei Wang, and Yongjun Xu. 2024. Relational Diffusion Distillation for Efficient Image Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) (*MM ’24*). Association for Computing Machinery, New York, NY, USA, 205–213. <https://doi.org/10.1145/3664647.3680768>
- [21] Azzedine Boukerche and Zhijun Hou. 2021. Object Detection Using Deep Learning Methods in Traffic Scenarios. *ACM Comput. Surv.* 54, 2, Article 30 (March 2021), 35 pages. <https://doi.org/10.1145/3434398>
- [22] Simone Antonelli, Danilo Avola, Luigi Cinque, Donato Crisostomi, Gian Luca Foresti, Fabio Galasso, Marco Raoul Marini, Alessio Mecca, and Daniele Pannone. 2022. Few-Shot Object Detection: A Survey. *ACM Comput. Surv.* 54, 11s, Article 242 (Sept. 2022), 37 pages. <https://doi.org/10.1145/3519022>
- [23] Jiaxu Leng, Yongming Ye, Mengjingcheng Mo, Chenqiang Gao, Ji Gan, Bin Xiao, and Xinbo Gao. 2024. Recent Advances for Aerial Object Detection: A Survey. *ACM Comput. Surv.* 56, 12, Article 296 (July 2024), 36 pages. <https://doi.org/10.1145/3664598>
- [24] Ayoub Benali Amjoud and Mustapha Amrouch. 2023. Object detection using deep learning, CNNs and vision transformers: A review. *IEEE Access* 11 (2023), 35479–35516.
- [25] Dingyuan Zhang, Dingkang Liang, Zhikang Zou, Jingyu Li, Xiaoqing Ye, Zhe Liu, Xiao Tan, and Xiang Bai. 2023. A simple vision transformer for weakly semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8373–8383.
- [26] Dahun Kim, Anelia Angelova, and Weicheng Kuo. 2023. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11144–11154.
- [27] Mathias Gehrig and Davide Scaramuzza. 2023. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13884–13893.
- [28] Tao Ye, Wenyang Qin, Zongyang Zhao, Xiaozhi Gao, Xiangpeng Deng, and Yu Ouyang. 2023. Real-time object detection network in UAV-vision based on CNN and transformer. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–13.
- [29] Sicheng Zhao, Huizai Yao, Chuang Lin, Yue Gao, and Guiguang Ding. 2024. Multi-source-free Domain Adaptive Object Detection. *International Journal of Computer Vision* (11 Jul 2024). <https://doi.org/10.1007/s11263-024-02170-z>
- [30] Xu Zhang, Zhe Chen, Jing Zhang, Tongliang Liu, and Dacheng Tao. 2024. Learning General and Specific Embedding with Transformer for Few-Shot Object Detection. *International Journal of Computer Vision* (Aug. 2024).
- [31] Zhe Chen, Jing Zhang, Yufei Xu, and Dacheng Tao. 2023. Transformer-Based Context Condensation for Boosting Feature Pyramids in Object Detection. *International Journal of Computer Vision* 131, 10 (Oct. 2023), 2738–2756.
- [32] Pengzhen Ren, Min Li, Zhen Luo, Xinshuai Song, Ziwei Chen, Weijia Liufu, Yixuan Yang, Hao Zheng, Rongtao Xu, Zitong Huang, et al. 2024. InfiniteWorld: A Unified Scalable Simulation Framework for General Visual-Language Robot Interaction. *arXiv preprint arXiv:2412.05789* (2024).
- [33] Rongtao Xu, Jian Zhang, Minghao Guo, Youpeng Wen, Haoting Yang, Min Lin, Jianzheng Huang, Zhe Li, Kaidong Zhang, Liqiong Wang, et al. 2025. A0: An Affordance-Aware Hierarchical Model for General Robotic Manipulation. *arXiv preprint arXiv:2504.12636* (2025).

- [34] Xiwen Liang, Min Lin, Weiqi Ruan, Rongtao Xu, Yuecheng Liu, Jiaqi Chen, Bingqian Lin, Yuzheng Zhuang, and Xiaodan Liang. 2025. Structured Preference Optimization for Vision-Language Long-Horizon Task Planning. *arXiv preprint arXiv:2502.20742* (2025).
- [35] Liang Ma, Jiajun Wen, Min Lin, Rongtao Xu, Xiwen Liang, Bingqian Lin, Jun Ma, Yongxin Wang, Ziming Wei, Haokun Lin, et al. 2025. PhyBlock: A Progressive Benchmark for Physical Understanding and Planning via 3D Block Assembly. *arXiv preprint arXiv:2506.08708* (2025).
- [36] Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, et al. 2025. Multimodal fusion and vision-language models: A survey for robot vision. *Information Fusion* (2025), 103652.
- [37] Zekai Zhang, Weiye Zhu, Hewei Pan, Xiangchen Wang, Rongtao Xu, Xing Sun, and Feng Zheng. 2025. ActiveVLN: Towards Active Exploration via Multi-Turn RL in Vision-and-Language Navigation. *arXiv preprint arXiv:2509.12618* (2025).
- [38] Shengli Zhou, Xiangchen Wang, Jinrui Zhang, Ruozai Tian, Rongtao Xu, and Feng Zheng. 2025.  $P^3$ : Toward Versatile Embodied Agents. *arXiv preprint arXiv:2508.07033* (2025).
- [39] Kaidong Zhang, Rongtao Xu, Pengzhen Ren, Junfan Lin, Hefeng Wu, Liang Lin, and Xiaodan Liang. 2025. RoBridge: A Hierarchical Architecture Bridging Cognition and Execution for General Robotic Manipulation. *arXiv preprint arXiv:2505.01709* (2025).
- [40] Rongtao Xu, Han Gao, Mingming Yu, Dong An, Shunpeng Chen, Changwei Wang, Li Guo, Xiaodan Liang, and Shibiao Xu. 2025. 3D-MoRe: Unified Modal-Contextual Reasoning for Embodied Question Answering. *arXiv preprint arXiv:2507.12026* (2025).
- [41] Kehan Chen, Dong An, Yan Huang, Rongtao Xu, Yifei Su, Yonggen Ling, Ian Reid, and Liang Wang. 2024. Constraint-Aware Zero-Shot Vision-Language Navigation in Continuous Environments. *arXiv preprint arXiv:2412.10137* (2024).
- [42] Yu Yan, Rongtao Xu, Jiazhao Zhang, Peiyang Li, Xiaodan Liang, and Jianqin Yin. 2024. InstruGen: Automatic Instruction Generation for Vision-and-Language Navigation Via Large Multimodal Models. *arXiv preprint arXiv:2411.11394* (2024).
- [43] Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and Wang He. 2024. NaVid: Video-based VLM Plans the Next Step for Vision-and-Language Navigation. *arXiv preprint arXiv:2402.15852* (2024).
- [44] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. 2023. Semantic segmentation using Vision Transformers: A survey. *Engineering Applications of Artificial Intelligence* 126 (2023), 106669.
- [45] Li Zhang, Jiachen Lu, Sixiao Zheng, Xinxuan Zhao, Xiatian Zhu, Yanwei Fu, Tao Xiang, Jianfeng Feng, and Philip H S Torr. 2024. Vision Transformers: From Semantic Segmentation to Dense Prediction. *International Journal of Computer Vision* (July 2024).
- [46] Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. 2023. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5240–5250.
- [47] Hengcan Shi, Munawar Hayat, and Jianfei Cai. 2023. Transformer scale gate for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3051–3060.
- [48] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. 2023. Semask: Semantically masked transformers for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 752–761.
- [49] Wenhao Xu, Rongtao Xu, Changwei Wang, Shibiao Xu, Li Guo, Man Zhang, and Xiaopeng Zhang. 2024. Spectral prompt tuning: Unveiling unseen classes for zero-shot semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6369–6377.
- [50] Wenhao Xu, Changwei Wang, Xuxiang Feng, Rongtao Xu, Longzhao Huang, Zherui Zhang, Li Guo, and Shibiao Xu. 2024. Generalization Boosted Adapter for Open-Vocabulary Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* (2024), 1–1. <https://doi.org/10.1109/TCSVT.2024.3454227>
- [51] Lianghui Zhu, Xinggang Wang, Jiapei Feng, Tianheng Cheng, Yingyue Li, Bo Jiang, Dingwen Zhang, and Junwei Han. 2024. WeakCLIP: Adapting CLIP for Weakly-Supervised Semantic Segmentation. *International Journal of Computer Vision* (05 Sep 2024). <https://doi.org/10.1007/s11263-024-02224-2>
- [52] Bowen Zhang, Liyang Liu, Minh Hieu Phan, Zhi Tian, Chunhua Shen, and Yifan Liu. 2024. SegViT v2: Exploring Efficient and Continual Semantic Segmentation with Plain Vision Transformers. *International Journal of Computer Vision* 132, 4 (April 2024), 1126–1147.
- [53] Swarnendu Ghosh, Nibaran Das, Ishita Das, and Ujjwal Maulik. 2019. Understanding Deep Learning Techniques for Image Segmentation. *ACM Comput. Surv.* 52, 4, Article 73 (Aug. 2019), 35 pages. <https://doi.org/10.1145/3329784>
- [54] Zherui Zhang, Changwei Wang, Rongtao Xu, Wenhao Xu, Shibiao Xu, Li Guo, Jiguang Zhang, Xiaoqiang Teng, and Wenbo Xu. 2024. MIM-HD: Making Smaller Masked Autoencoder Better with Efficient Distillation. In *European Conference on Artificial Intelligence*. <https://api.semanticscholar.org/CorpusID:273589355>
- [55] Rongtao Xu, Changwei Wang, Jiguang Zhang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. 2024. SkinFormer: Learning Statistical Texture Representation With Transformer for Skin Lesion Segmentation. *IEEE Journal of Biomedical and Health Informatics* 28 (2024), 6008–6018. <https://api.semanticscholar.org/CorpusID:270710189>
- [56] Rongtao Xu, Changwei Wang, Duzhen Zhang, Man Zhang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. 2024. DeffFusion: Deformable Multimodal Representation Fusion for 3D Semantic Segmentation. *2024 IEEE International Conference on Robotics and Automation (ICRA)* (2024), 7732–7739. <https://api.semanticscholar.org/CorpusID:271798187>
- [57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
- [58] Hamid Tabani, Ajay Balasubramaniam, Shabbir Marzban, Elahe Arani, and Bahram Zonooz. 2021. Improving the efficiency of transformers for resource-constrained devices. In *2021 24th Euromicro Conference on Digital System Design (DSD)*. IEEE, 449–456.
- [59] Feiyang Chen, Ziqian Luo, Lisang Zhou, Xueting Pan, and Ying Jiang. 2024. Comprehensive Survey of Model Compression and Speed up for Vision Transformers. *arXiv:2404.10407* [cs.CV] <https://arxiv.org/abs/2404.10407>

- [60] Hou-I Liu, Marco Galindo, Hongxia Xie, Lai-Kuan Wong, Hong-Han Shuai, Yung-Hui Li, and Wen-Huang Cheng. 2024. Lightweight Deep Learning for Resource-Constrained Environments: A Survey. *ACM Comput. Surv.* 56, 10, Article 267 (June 2024), 42 pages. <https://doi.org/10.1145/3657282>
- [61] Quentin Fournier, Gaétan Marceau Caron, and Daniel Aloise. 2023. A Practical Survey on Faster and Lighter Transformers. *ACM Comput. Surv.* 55, 14s, Article 304 (July 2023), 40 pages. <https://doi.org/10.1145/3586074>
- [62] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient Transformers: A Survey. *ACM Comput. Surv.* 55, 6, Article 109 (Dec. 2022), 28 pages. <https://doi.org/10.1145/3530811>
- [63] Lu Yu and Wei Xiang. 2023. X-pruner: explainable pruning for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 24355–24363.
- [64] Lei Liu, Gary G Yen, and Zhenan He. 2024. EvolutionViT: Multi-objective evolutionary vision transformer pruning under resource constraints. *Information Sciences* (2024), 121406.
- [65] Hao Yu and Jianxin Wu. 2023. A unified pruning framework for vision transformers. *Science China Information Sciences* 66, 7 (2023), 179101.
- [66] Chong Yu, Tao Chen, Zhongxue Gan, and Jiayuan Fan. 2023. Boost vision transformer with gpu-friendly sparsity and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22658–22668.
- [67] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. 2023. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17227–17236.
- [68] Dayou Du, Gu Gong, and Xiaowen Chu. 2024. Model Quantization and Hardware Acceleration for Vision Transformers: A Comprehensive Survey. arXiv:2405.00314 [cs.LG] <https://arxiv.org/abs/2405.00314>
- [69] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. 2023. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5961–5971.
- [70] Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. 2023. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14420–14430.
- [71] Borui Zhao, Renjie Song, and Jiajun Liang. 2023. Cumulative spatial knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6146–6155.
- [72] Yuliang Cai, Jesse Thomason, and Mohammad Rostami. 2023. Task-attentive transformer architecture for continual learning of vision-and-language tasks using knowledge distillation. *arXiv preprint arXiv:2303.14423* (2023).
- [73] Zherui Zhang, Changwei Wang, Rongtao Xu, Wenhao Xu, Shibiao Xu, Yu Zhang, and Li Guo. 2025. CAE-DFKD: Bridging the Transferability Gap in Data-Free Knowledge Distillation. <https://api.semanticscholar.org/CorpusID:278207768>
- [74] Zherui Zhang, Rongtao Xu, Changwei Wang, Wenhao Xu, Shunpeng Chen, Shibiao Xu, Guangyuan Xu, and Li Guo. 2025. DFMC:Feature-Driven Data-Free Knowledge Distillation. *IEEE Transactions on Circuits and Systems for Video Technology* (2025), 1–1. <https://doi.org/10.1109/TCSVT.2025.3565616>
- [75] Yongming Rao, Zuyan Liu, Wenliang Zhao, Jie Zhou, and Jiwen Lu. 2023. Dynamic spatial sparsification for efficient vision transformers and convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [76] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. 2023. Dat++: Spatially dynamic vision transformer with deformable attention. *arXiv preprint arXiv:2309.01430* (2023).
- [77] Xin Lai, Yuhui Yuan, Ruihang Chu, Yukang Chen, Han Hu, and Jiaya Jia. 2023. Mask-attention-free transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3693–3703.
- [78] Rina Buoy, Masakazu Iwamura, Sovila Sruri, and Koichi Kise. 2023. ViTSTR-Transducer: Cross-Attention-Free Vision Transformer Transducer for Scene Text Recognition. *Journal of Imaging* 9, 12 (2023), 276.
- [79] Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. 2022. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12020–12030.
- [80] Kai Zhang, Zhuolin Li, Feng Zhang, Wenbo Wan, and Jiande Sun. 2022. Pan-sharpening based on transformer with redundancy reduction. *IEEE Geoscience and Remote Sensing Letters* 19 (2022), 1–5.
- [81] Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. 2021. IA-RED<sup>2</sup>: Interpretability-Aware Redundancy Reduction for Vision Transformers. *Advances in Neural Information Processing Systems* 34 (2021), 24898–24911.
- [82] Zhanzhou Feng and Shiliang Zhang. 2023. Efficient vision transformer via token merger. *IEEE Transactions on Image Processing* (2023).
- [83] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. 2024. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1383–1392.
- [84] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. arXiv:2103.14899 [cs.CV]
- [85] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. 2021. CvT: Introducing Convolutions to Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22–31.
- [86] Sachin Mehta and Mohammad Rastegari. 2022. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. arXiv:2110.02178 [cs.CV] <https://arxiv.org/abs/2110.02178>
- [87] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. 2022. Mobile-Former: Bridging MobileNet and Transformer. arXiv:2108.05895 [cs.CV]

- [88] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. 2022. Shunted Self-Attention via Multi-Scale Token Aggregation. arXiv:2111.15193 [cs.CV]
- [89] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. 2022. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12165–12174.
- [90] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. RepViT: Revisiting Mobile CNN From ViT Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15909–15920.
- [91] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. 2022. TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation. arXiv:2204.05525 [cs.CV]
- [92] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-Token ViT: Training Vision Transformers From Scratch on ImageNet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 558–567.
- [93] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems* 34 (2021), 13937–13949.
- [94] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. 2023. Conditional Positional Encodings for Vision Transformers. arXiv:2102.10882 [cs.CV]
- [95] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. 2024. Rotary Position Embedding for Vision Transformer. arXiv:2403.13298 [cs.CV]
- [96] Zhicai Wang, Yanbin Hao, Xingyu Gao, Hao Zhang, Shuo Wang, Tingting Mu, and Xiangnan He. 2022. Parameterization of Cross-token Relations with Relative Positional Encoding for Vision MLP. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. ACM. <https://doi.org/10.1145/3503161.3547953>
- [97] Heeseung Kwon, Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, and Kartek Alahari. 2022. Lightweight Structure-Aware Attention for Visual Understanding. arXiv:2211.16289 [cs.CV]
- [98] Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. 2024. Functional Interpolation for Relative Positions improves Long Context Transformers. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=rR03qFesqk>
- [99] Zhen Qin, Weixuan Sun, Kaiyue Lu, Hui Deng, Dongxu Li, XiaoDong Han, Yuchao Dai, Lingpeng Kong, and Yiran Zhong. 2023. Relative Positional Encoding Family via Unitary Transformation. <https://openreview.net/forum?id=xMWFqb5Uyk>
- [100] Mohamed Yacin Sikkandar, Sankar Ganesh Sundaram, Ahmad Allassaf, Ibrahim AlMohimeed, Khalid Alhussaini, Adham Aleid, Salem Ali Alolayan, P Ramkumar, Meshal Khalaf Almutairi, and Sabarunisha Begum. 2024. Utilizing adaptive deformable convolution and position embedding for colon polyp segmentation with a visual transformer. *Scientific reports* 14 1 (2024), 7318. <https://api.semanticscholar.org/CorpusID:268731979>
- [101] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. 2022. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. arXiv:2107.00652 [cs.CV]
- [102] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. 2021. Global Filter Networks for Image Classification. arXiv:2107.00645 [cs.CV]
- [103] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. 2021. Focal Self-attention for Local-Global Interactions in Vision Transformers. arXiv:2107.00641 [cs.CV]
- [104] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. 2022. MaxViT: Multi-Axis Vision Transformer. arXiv:2204.01697 [cs.CV]
- [105] Shashanka Venkataraman, Amir Ghodrati, Yuki M. Asano, Fatih Porikli, and Amirsossein Habibian. 2023. Skip-Attention: Improving Vision Transformers by Paying Less Attention. arXiv:2301.02240 [cs.CV]
- [106] Abdelrahman Shaker, Muhammad Maaz, Hanooona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023. SwiftFormer: Efficient Additive Attention for Transformer-based Real-time Mobile Vision Applications. arXiv:2303.15446 [cs.CV]
- [107] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson Lau. 2023. BiFormer: Vision Transformer with Bi-Level Routing Attention. arXiv:2303.08810 [cs.CV]
- [108] Anxhelo Diko, Danilo Avola, Marco Cascio, and Luigi Cinque. 2024. ReViT: Enhancing Vision Transformers with Attention Residual Connections for Visual Recognition. arXiv:2402.11301 [cs.CV]
- [109] Qihang Fan, Huabo Huang, Mingrui Chen, and Ran He. 2024. Vision Transformer with Sparse Scan Prior. arXiv:2405.13335 [cs.CV]. <https://arxiv.org/abs/2405.13335>
- [110] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 [cs.CV]
- [111] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. 2021. Transformer in Transformer. In *Neural Information Processing Systems*. <https://api.semanticscholar.org/CorpusID:232076027>
- [112] Jingkai Zhou, Pichao Wang, Jiasheng Tang, Fan Wang, Qiong Liu, Hao Li, and Rong Jin. 2023. What Limits the Performance of Local Self-attention? *International Journal of Computer Vision* 131, 10 (Oct. 2023), 2516–2528.
- [113] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. 2022. Vision Transformer with Deformable Attention. arXiv:2201.00520 [cs.CV]
- [114] Masato Tamura. 2024. Design and Analysis of Efficient Attention in Transformers for Social Group Activity Recognition. *International Journal of Computer Vision* (May 2024).

- [115] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. 2021. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2998–3008.
- [116] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. MLP-Mixer: An all-MLP Architecture for Vision. arXiv:2105.01601 [cs.CV]
- [117] Jiachen Lu, Junge Zhang, Xiatian Zhu, Jianfeng Feng, Tao Xiang, and Li Zhang. 2024. Softmax-Free Linear Transformers. *International Journal of Computer Vision* 132, 8 (Aug. 2024), 3355–3374.
- [118] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. arXiv:2001.04451 [cs.LG]
- [119] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-Attention with Linear Complexity. arXiv:2006.04768 [cs.LG]
- [120] Sachin Mehta and Mohammad Rastegari. 2022. Separable Self-attention for Mobile Vision Transformers. arXiv:2206.02680 [cs.CV]
- [121] Weixuan Sun, Zhen Qin, Hui Deng, Jianyuan Wang, Yi Zhang, Kaihao Zhang, Nick Barnes, Stan Birchfield, Lingpeng Kong, and Yiran Zhong. 2023. Vicinity Vision Transformer. arXiv:2206.10552 [cs.CV]
- [122] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. 2023. VOLO: Vision Outlooker for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 5 (2023), 6575–6586. <https://doi.org/10.1109/TPAMI.2022.3206108>
- [123] Sachin Mehta and Mohammad Rastegari. 2021. MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer. *arXiv preprint arXiv:2110.02178* (2021).
- [124] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Francois Fleuret. 2020. Faster Attention Is What You Need: Fast self-attention with linear complexity. *arXiv preprint arXiv:2009.14794* (2020).
- [125] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2020. Rethinking Attention with Performers. *arXiv preprint arXiv:2009.14794* (2020).
- [126] Alexander Novikov, Mikhail Trofimov, and Ivan Oseledets. 2020. Neural network factorization and inversion using tensor decomposition. *arXiv preprint arXiv:2009.02523* (2020).
- [127] Vadim Lebedev, Yaroslav Ganin, Frank Rudzicz, and Victor Lempitsky. 2014. Speeding-up Convolutional Neural Networks Using Fine-tuned CP-decomposition. In *Advances in Neural Information Processing Systems*. 1780–1788.
- [128] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2021. Deep Equilibrium Models. *Advances in Neural Information Processing Systems* 33 (2021), 9797–9812.
- [129] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. 2021. ResMLP: Feedforward networks for image classification with data-efficient training. arXiv:2105.03404 [cs.CV]
- [130] Tan Yu, Xu Li, Yunfeng Cai, Mingming Sun, and Ping Li. 2021. S<sup>2</sup>-MLP: Spatial-Shift MLP Architecture for Vision. arXiv:2106.07477 [cs.CV]
- [131] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. 2022. AS-MLP: An Axial Shifted MLP Architecture for Vision. arXiv:2107.08391 [cs.CV]
- [132] Franchis N. Saikia, Yuji Iwahori, Taisei Suzuki, M. K. Bhuyan, Aili Wang, and Boonserm Kijisirikul. 2023. MLP-UNet: Glomerulus Segmentation. *IEEE Access* 11 (2023), 53034–53047. <https://doi.org/10.1109/ACCESS.2023.3280831>
- [133] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. 2021. Pay attention to mlps. *Advances in neural information processing systems* 34 (2021), 9204–9215.
- [134] Jianyuan Guo, Yehui Tang, Kai Han, Xinghao Chen, Han Wu, Chao Xu, Chang Xu, and Yunhe Wang. 2021. Hire-MLP: Vision MLP via Hierarchical Rearrangement. arXiv:2108.13341 [cs.CV]
- [135] Jiachen Li, Ali Hassani, Steven Walton, and Humphrey Shi. 2021. ConvMLP: Hierarchical Convolutional MLPs for Vision. arXiv:2109.04454 [cs.CV]
- [136] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. 2022. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE transactions on pattern analysis and machine intelligence* 45, 4 (2022), 5314–5321.
- [137] Shoufa Chen, Enze Xie, Chongjian Ge, Runjian Chen, Ding Liang, and Ping Luo. 2022. CycleMLP: A MLP-like Architecture for Dense Prediction. arXiv:2107.10224 [cs.CV]
- [138] Tan Yu, Xu Li, Yunfeng Cai, Mingming Sun, and Ping Li. 2022. S2-mlp: Spatial-shift mlp architecture for vision. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 297–306.
- [139] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. 2022. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10819–10829.
- [140] Badri Patro and Vijay Agnieszwaran. 2024. Scattering vision transformer: Spectral mixing matters. *Advances in Neural Information Processing Systems* 36 (2024).
- [141] Yixin Cheng, Grigoris G. Chrysos, Markos Georgopoulos, and Volkan Cevher. 2024. Multilinear Operator Networks. arXiv:2401.17992 [cs.CV] <https://arxiv.org/abs/2401.17992>
- [142] Xinyue Wang, Zhicheng Cai, and Chenglei Peng. 2023. X-MLP: A Patch Embedding-Free MLP Architecture for Vision. arXiv:2307.00592 [cs.CV]
- [143] Junxiong Wang, Jing Nathan Yan, Albert Gu, and Alexander M. Rush. 2023. Pretraining Without Attention. arXiv:2212.10544 [cs.CL] <https://arxiv.org/abs/2212.10544>

- [144] Chao Ji, Zhaohong Deng, Yan Ding, Fengsheng Zhou, and Zhiyong Xiao. 2023. RMMILP:Rolling MLP and matrix decomposition for skin lesion segmentation. *Biomed. Signal Process. Control.* 84 (2023), 104825. <https://api.semanticscholar.org/CorpusID:257439689>
- [145] Miao Cao, Lishun Wang, Mingyu Zhu, and Xin Yuan. 2024. Hybrid CNN-Transformer Architecture for Efficient Large-Scale Video Snapshot Compressive Imaging. *International Journal of Computer Vision* (May 2024).
- [146] Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently Modeling Long Sequences with Structured State Spaces. arXiv:2111.00396 [cs.LG] <https://arxiv.org/abs/2111.00396>
- [147] Jiaming Luo, Yongzhen Tang, Jie Wang, and Hongtao Lu. 2023. USMLP: U-shaped Sparse-MLP network for mass segmentation in mammograms. *Image Vis. Comput.* 137 (2023), 104761. <https://api.semanticscholar.org/CorpusID:259614161>
- [148] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. 2022. Long Range Language Modeling via Gated State Spaces. arXiv:2206.13947 [cs.LG] <https://arxiv.org/abs/2206.13947>
- [149] Albert Gu and Tri Dao. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752 [cs.LG] <https://arxiv.org/abs/2312.00752>
- [150] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. arXiv:2401.09417 [cs.CV] <https://arxiv.org/abs/2401.09417>
- [151] Jing Yao, Danfeng Hong, Chenyu Li, and Jocelyn Chanussot. 2024. SpectralMamba: Efficient Mamba for Hyperspectral Image Classification. arXiv:2404.08489 [cs.CV] <https://arxiv.org/abs/2404.08489>
- [152] Xiaohuan Pei, Tao Huang, and Chang Xu. 2024. EfficientVMamba: Atrous Selective Scan for Light Weight Visual Mamba. arXiv:2403.09977 [cs.CV] <https://arxiv.org/abs/2403.09977>
- [153] Tianxiang Chen, Zi Ye, Zhentao Tan, Tao Gong, Yue Wu, Qi Chu, Bin Liu, Nenghai Yu, and Jieping Ye. 2024. MiM-ISTD: Mamba-in-Mamba for Efficient Infrared Small Target Detection. arXiv:2403.02148 [cs.CV] <https://arxiv.org/abs/2403.02148>
- [154] Yujin Tang, Peijie Dong, Zhenheng Tang, Xiaowen Chu, and Junwei Liang. 2024. VMRNN: Integrating Vision Mamba and LSTM for Efficient and Accurate Spatiotemporal Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 5663–5673.
- [155] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jie Zhou, and Jiwen Lu. 2023. GFNet: Global filter networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 10960–10973.
- [156] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*. 568–578.
- [157] Yun Liu, Yu-Huan Wu, Guolei Sun, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. 2024. Vision Transformers with Hierarchical Attention. arXiv:2106.03180 [cs.CV] <https://arxiv.org/abs/2106.03180>
- [158] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. 2021. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. arXiv:2104.13840 [cs.CV]
- [159] Yixing Xu, Chao Li, Dong Li, Xiao Sheng, Fan Jiang, Lu Tian, and Ashish Sirasao. 2023. FDViT: Improve the Hierarchical Architecture of Vision Transformer. 5927–5937. <https://doi.org/10.1109/ICCV51070.2023.00547>
- [160] Yixing Xu, Chao Li, Dong Li, Xiao Sheng, Fan Jiang, Lu Tian, and Ashish Sirasao. 2023. Fdvit: Improve the hierarchical architecture of vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5950–5960.
- [161] Xiaosong Zhang, Yunjie Tian, Lingxi Xie, Wei Huang, Qi Dai, Qixiang Ye, and Qi Tian. 2023. Hivit: A simpler and more efficient design of hierarchical vision transformer. In *The Eleventh International Conference on Learning Representations*.
- [162] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. 2023. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. arXiv:2306.00989 [cs.CV]
- [163] Ting Yao, Yehao Li, Yingwei Pan, and Tao Mei. 2024. HIRI-ViT: Scaling Vision Transformer With High Resolution Inputs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 9 (2024), 6431–6442. <https://doi.org/10.1109/TPAMI.2024.3379457>
- [164] Shuzhe Wang, Zakaria Laskar, Iaroslav Melekhov, Xiaotian Li, Yi Zhao, Giorgos Tolias, and Juho Kannala. 2024. HSCNet++: Hierarchical Scene Coordinate Classification and Regression for Visual Localization with Transformer. *International Journal of Computer Vision* 132, 7 (July 2024), 2530–2550.
- [165] Yun Liu, Yu-Huan Wu, Guolei Sun, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. 2024. Vision transformers with hierarchical attention. *Machine Intelligence Research* (2024), 1–14.
- [166] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. 2021. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11936–11945.
- [167] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. 2021. Benchmarking Detection Transfer Learning with Vision Transformers. arXiv:2111.11429 [cs.CV]
- [168] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. 2023. Vision Transformer Adapter for Dense Predictions. arXiv:2205.08534 [cs.CV]
- [169] Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, and Hongsheng Li. 2023. MixMAE: Mixed and Masked Autoencoder for Efficient Pretraining of Hierarchical Vision Transformers. arXiv:2205.13137 [cs.CV]

- [170] Xiaosong Zhang, Yunjie Tian, Wei Huang, Qixiang Ye, Qi Dai, Lingxi Xie, and Qi Tian. 2022. HiViT: Hierarchical Vision Transformer Meets Masked Image Modeling. arXiv:2205.14949 [cs.CV]
- [171] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. 2022. Green Hierarchical Vision Transformer for Masked Image Modeling. arXiv:2205.13515 [cs.CV]
- [172] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. 2021. Evo-ViT: Slow-Fast Token Evolution for Dynamic Vision Transformer. arXiv:2108.01390 [cs.CV]
- [173] Joyce Zheng, Mehdi Rezagholizadeh, and Peyman Passban. 2022. Dynamic Position Encoding for Transformers. arXiv:2204.08142 [cs.CL]
- [174] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. 2022. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. arXiv:2202.07800 [cs.CV]
- [175] Zhenglun Kong, Peiyang Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. 2022. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *European Conference on Computer Vision*. Springer, 620–640.
- [176] Peiyang Dong, Mengshu Sun, Alec Lu, Yanyue Xie, Kenneth Liu, Zhenglun Kong, Xin Meng, Zhengang Li, Xue Lin, Zhenman Fang, and Yanzhi Wang. 2023. HeatViT: Hardware-Efficient Adaptive Token Pruning for Vision Transformers. In *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. 442–455. <https://doi.org/10.1109/HPCA56546.2023.10071047>
- [177] Mengzhao Chen, Mingbao Lin, Ke Li, Yunhang Shen, Yongjian Wu, Fei Chao, and Rongrong Ji. 2022. CF-ViT: A General Coarse-to-Fine Method for Vision Transformer. arXiv:2203.03821 [cs.CV]
- [178] Xuwei Xu, Changlin Li, Yudong Chen, Xiaojun Chang, Jiajun Liu, and Sen Wang. 2023. No Token Left Behind: Efficient Vision Transformer via Dynamic Token Idling. arXiv:2310.05654 [cs.CV]
- [179] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Jozé, Eric Sommerladé, Hamed Pirsiavash, and Jürgen Gall. 2022. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*. Springer, 396–414.
- [180] Wentao Zhu. 2024. TPC-ViT: Token Propagation Controller for Efficient Vision Transformer. arXiv:2401.01470 [cs.CV]
- [181] Zhe Bian, Zhe Wang, Wenqiang Han, and Kangping Wang. 2023. Multi-Scale And Token Mergence: Make Your ViT More Efficient. arXiv:2306.04897 [cs.CV]
- [182] Mingbao Lin, Mengzhao Chen, Yuxin Zhang, Chunhua Shen, Rongrong Ji, and Liujuan Cao. 2023. Super Vision Transformer. arXiv:2205.11397 [cs.CV]
- [183] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token Merging: Your ViT But Faster. arXiv:2210.09461 [cs.CV]
- [184] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. 2023. Token Fusion: Bridging the Gap between Token Pruning and Token Merging. arXiv:2312.01026 [cs.CV]
- [185] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. 2022. Not All Tokens Are Equal: Human-centric Visual Analysis via Token Clustering Transformer. arXiv:2204.08680 [cs.CV]
- [186] Meiqi Wang, Zhisheng Wang, Jinming Lu, Jun Lin, and Zhongfeng Wang. 2019. E-LSTM: An Efficient Hardware Architecture for Long Short-Term Memory. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 2 (2019), 280–291. <https://doi.org/10.1109/JETCAS.2019.2911739>
- [187] Boyu Chen, Peixia Li, Baopu Li, Chuming Li, Lei Bai, Chen Lin, Ming Sun, Junjie Yan, and Wanli Ouyang. 2021. PSViT: Better Vision Transformer via Token Pooling and Attention Sharing. arXiv:2108.03428 [cs.CV] <https://arxiv.org/abs/2108.03428>
- [188] Shibo Jie, Yehui Tang, Jianyuan Guo, Zhi-Hong Deng, Kai Han, and Yunhe Wang. 2024. Token Compensator: Altering Inference Cost of Vision Transformer without Re-Tuning. <https://api.semanticscholar.org/CorpusID:271860275>
- [189] Hao Wang, Wenjia Zhang, and Guohua Liu. 2023. TSNet: Token Sparsification for Efficient Video Transformer. *Applied Sciences* 13, 19 (2023). <https://doi.org/10.3390/app131910633>
- [190] Cedric Renggli, André Susano Pinto, Neil Houlsby, Basil Mustafa, Joan Puigcerver, and Carlos Riquelme. 2022. Learning to Merge Tokens in Vision Transformers. arXiv:2202.12015 [cs.CV]
- [191] Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, and Jiajun Liang. 2023. Joint Token Pruning and Squeezing Towards More Aggressive Compression of Vision Transformers. arXiv:2304.10716 [cs.CV]
- [192] Zhanzhou Feng and Shiliang Zhang. 2023. Efficient Vision Transformer via Token Merger. *IEEE Transactions on Image Processing* 32 (2023), 4156–4169. <https://doi.org/10.1109/TIP.2023.3293763>
- [193] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. *CoRR* abs/2101.11986 (2021). arXiv:2101.11986 <https://arxiv.org/abs/2101.11986>
- [194] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. 2022. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. In *International Conference on Learning Representations*. [https://openreview.net/forum?id=BjyvwnXVn\\_](https://openreview.net/forum?id=BjyvwnXVn_)
- [195] Hongxu Yin, Arash Vahdat, Jose M. Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. 2022. A-ViT: Adaptive Tokens for Efficient Vision Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10809–10818.
- [196] Mary Phuong and Christoph Lampert. 2019. Distillation-Based Training for Multi-Exit Architectures. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 1355–1364. <https://doi.org/10.1109/ICCV.2019.00144>
- [197] Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. 2022. Single-Layer Vision Transformers for More Accurate Early Exits with Less Overhead. arXiv:2105.09121 [cs.LG]
- [198] Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. 2021. Multi-Exit Vision Transformer for Dynamic Inference. arXiv:2106.15183 [cs.CV]

- [199] Guanyu Xu, Jiawei Hao, Li Shen, Han Hu, Yong Luo, Hui Lin, and Jiale Shen. 2023. LGViT: Dynamic Early Exiting for Accelerating Vision Transformer. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. ACM. <https://doi.org/10.1145/3581783.3611762>
- [200] Yurong Zhang, Honghao Chen, Xinyu Zhang, Xiangxiang Chu, and Li Song. 2024. Dyn-Adapter: Towards Disentangled Representation for Efficient Visual Recognition. arXiv:2407.14302 [cs.CV] <https://arxiv.org/abs/2407.14302>
- [201] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser Nam Lim. 2021. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021), 12299–12308. <https://api.semanticscholar.org/CorpusID:244729636>
- [202] Yijin Liu, Fandong Meng, Jie Zhou, Yufeng Chen, and Jinan Xu. 2021. Faster depth-adaptive transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13424–13432.
- [203] Mahdi Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. 2020. Depth-Adaptive Transformer. arXiv:1910.10073 [cs.CL] <https://arxiv.org/abs/1910.10073>
- [204] Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems* 35 (2022), 17456–17472.
- [205] Luciano Del Corro, Allie Del Giorno, Sahaj Agarwal, Bin Yu, Ahmed Awadallah, and Subhabrata Mukherjee. 2023. SkipDecode: Autoregressive Skip Decoding with Batching and Caching for Efficient LLM Inference. arXiv:2307.02628 [cs.CL] <https://arxiv.org/abs/2307.02628>
- [206] Mengzhao Chen, Wenqi Shao, Peng Xu, Mingbao Lin, Kaipeng Zhang, Fei Chao, Rongrong Ji, Yu Qiao, and Ping Luo. 2023. DiffRate : Differentiable Compression Rate for Efficient Vision Transformers. arXiv:2305.17997 [cs.CV] <https://arxiv.org/abs/2305.17997>
- [207] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [stat.ML] <https://arxiv.org/abs/1503.02531>
- [208] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge Distillation: A Survey. *International Journal of Computer Vision* 129, 6 (June 2021), 1789–1819.
- [209] Kai Wang, Fei Yang, and Joost van de Weijer. 2022. Attention Distillation: self-supervised vision transformer students need more guidance. arXiv:2210.00944 [cs.CV] <https://arxiv.org/abs/2210.00944>
- [210] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. 2022. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 1306–1313.
- [211] Chunyu Fan, Qi Su, Zhifeng Xiao, Hao Su, Aijie Hou, and Bo Luan. 2023. ViT-FRD: A vision transformer model for cardiac MRI image segmentation based on feature recombination distillation. *IEEE Access* (2023).
- [212] Wujie Sun, Defang Chen, Can Wang, Deshi Ye, Yan Feng, and Chun Chen. 2023. Accelerating diffusion sampling with classifier-based feature distillation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 810–815.
- [213] Minki Ji, Byeongho Heo, and Sungrae Park. 2021. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7945–7952.
- [214] Benjamin Kelenyi, Victor Domsa, and Levente Tamas. 2024. SAM-Net: self-attention based feature matching with spatial transformers and knowledge distillation. *Expert Systems with Applications* 242 (2024), 122804.
- [215] Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. 2024. ViTKD: Feature-based Knowledge Distillation for Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1379–1388.
- [216] Jun Zhang, Yunfei Zhang, Feixue Shao, Xuetao Ma, and Daoxiang Zhou. 2023. KDFAS: Multi-stage Knowledge Distillation Vision Transformer for Face Anti-spoofing. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 159–171.
- [217] Guang Yang, Yin Tang, Zhijian Wu, Jun Li, Jianhua Xu, and Xili Wan. 2024. DMKD: Improving Feature-Based Knowledge Distillation for Object Detection Via Dual Masking Augmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3330–3334.
- [218] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. 2024. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26617–26626.
- [219] Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. 2024. CrossKD: Cross-head knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16520–16530.
- [220] Jianping Gou, Xiangshuo Xiong, Baosheng Yu, Lan Du, Yibing Zhan, and Dacheng Tao. 2023. Multi-target knowledge distillation via student self-reflection. *International Journal of Computer Vision* 131, 7 (2023), 1857–1874.
- [221] Yaohui Zhu, Linhu Liu, and Jiang Tian. 2023. Learn more for food recognition via progressive self-distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 3879–3887.
- [222] Conghao Xiong, Yi-Mou Lin, Hao Chen, Joseph J. Y. Sung, and Irwin King. 2023. TAKT: Target-Aware Knowledge Transfer for Whole Slide Image Classification. <https://api.semanticscholar.org/CorpusID:257482790>
- [223] Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, Jingting Ding, and Liang Li. 2022. Cross-architecture knowledge distillation. In *Proceedings of the Asian conference on computer vision*. 3396–3411.
- [224] Yao Ni, Piotr Koniusz, Richard Hartley, and Richard Nock. 2022. Manifold learning benefits GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11265–11274.
- [225] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. 2022. Dearkd: data-efficient early knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12052–12062.

- [226] Peng Gao, Ziyi Lin, Renrui Zhang, Rongyao Fang, Hongyang Li, Hongsheng Li, and Yu Qiao. 2024. Mimic before Reconstruct: Enhancing Masked Autoencoders with Feature Mimicking. *International Journal of Computer Vision* 132, 5 (May 2024), 1546–1556.
- [227] Hailin Zhang, Defang Chen, and Can Wang. 2023. Adaptive multi-teacher knowledge distillation with meta-learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1943–1948.
- [228] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. 2022. Knowledge distillation via the target-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10915–10924.
- [229] Sucheng Ren, Fangyun Wei, Zheng Zhang, and Han Hu. 2023. Tinymim: An empirical study of distilling mim pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3687–3697.
- [230] Liang Zhao, Yao Teng, and Limin Wang. 2024. Logit Normalization for Long-Tail Object Detection. *International Journal of Computer Vision* 132, 6 (June 2024), 2114–2134.
- [231] Jia Guo. 2022. Reducing the Teacher-Student Gap via Adaptive Temperatures. [https://openreview.net/forum?id=h-z\\_zqT2yJU](https://openreview.net/forum?id=h-z_zqT2yJU)
- [232] Jihao Liu, Boxiao Liu, Hongsheng Li, and Yu Liu. 2022. Meta Knowledge Distillation. arXiv:2202.07940 [cs.LG] <https://arxiv.org/abs/2202.07940>
- [233] Ying Jin, Jiaqi Wang, and Dahu Lin. 2023. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24276–24285.
- [234] Zhihao Chi, Tu Zheng, Hengjia Li, Zheng Yang, Boxi Wu, Binbin Lin, and Deng Cai. 2023. NormKD: Normalized Logits for Knowledge Distillation. arXiv:2308.00520 [cs.CV] <https://arxiv.org/abs/2308.00520>
- [235] Mengyang Yuan, Bo Lang, and Fengnan Quan. 2024. Student-friendly knowledge distillation. *Knowledge-Based Systems* 296 (2024), 111915.
- [236] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. 2024. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15731–15740.
- [237] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 11953–11962.
- [238] Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. 2023. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17185–17194.
- [239] Rongzhi Zhang, Jiaming Shen, Tianqi Liu, Jialu Liu, Michael Bendersky, Marc Najork, and Chao Zhang. 2023. Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. *arXiv preprint arXiv:2305.05010* (2023).
- [240] Chuan Li, Xiao Teng, Yan Ding, and Long Lan. 2024. NTCE-KD: Non-Target-Class-Enhanced Knowledge Distillation. *Sensors* 24, 11 (2024), 3617.
- [241] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. 2020. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3903–3911.
- [242] Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. 2024. One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation. *Advances in Neural Information Processing Systems* 36 (2024).
- [243] Jing Ma, Xiang Xiang, Ke Wang, Yuchuan Wu, and Yongbin Li. 2024. Aligning Logits Generatively for Principled Black-Box Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23148–23157.
- [244] Kaixiang Zheng and En-Hui Yang. 2024. Knowledge Distillation Based on Transformed Teacher Matching. arXiv:2402.11148 [cs.LG] <https://arxiv.org/abs/2402.11148>
- [245] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. 2022. Tinyvit: Fast pretraining distillation for small vision transformers. In *European conference on computer vision*. Springer, 68–85.
- [246] Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, Jingting Ding, Liang Li, and Stephen Maybank. 2024. Cross-Architecture Knowledge Distillation. *International Journal of Computer Vision* 132, 8 (Aug. 2024), 2798–2824.
- [247] Shicai Wei, Chunbo Luo, and Yang Luo. 2024. Scaled Decoupled Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15975–15983.
- [248] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [249] Qianlong Xiang, Miao Zhang, Yuzhang Shang, Jianlong Wu, Yan Yan, and Liqiang Nie. 2024. DKDM: Data-Free Knowledge Distillation for Diffusion Models with Any Architecture. arXiv:2409.03550 [cs.CV] <https://arxiv.org/abs/2409.03550>
- [250] Minh-Tuan Tran, Trung Le, Xuan-May Le, Mehrtash Harandi, Quan Hung Tran, and Dinh Phung. 2024. Nayer: Noisy layer data generation for efficient and effective data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23860–23869.