
DropKey

Bonan Li*

Yinhan Hu

Xuecheng Nie

Congying Han

Xiangjian Jiang

Tiande Guo

Luoqi Liu

Abstract

In this paper, we focus on analyzing and improving the dropout technique for self-attention layers of Vision Transformer, which is important while surprisingly ignored by prior works. In particular, we conduct researches on three core questions: First, *what to drop in self-attention layers?* Different from dropping attention weights in literature, we propose to move dropout operations forward ahead of attention matrix calculation and set the Key as the dropout unit, yielding a novel dropout-before-softmax scheme. We theoretically verify that this scheme helps keep both regularization and probability features of attention weights, alleviating the overfittings problem to specific patterns and enhancing the model to globally capture vital information; Second, *how to schedule the drop ratio in consecutive layers?* In contrast to exploit a constant drop ratio for all layers, we present a new decreasing schedule that gradually decreases the drop ratio along the stack of self-attention layers. We experimentally validate the proposed schedule can avoid overfittings in low-level features and missing in high-level semantics, thus improving the robustness and stableness of model training; Third, *whether need to perform structured dropout operation as CNN?* We attempt patch-based block-version of dropout operation and find that this useful trick for CNN is not essential for ViT. Given exploration on the above three questions, we present the novel DropKey method that regards Key as the drop unit and exploits decreasing schedule for drop ratio, improving ViTs in a general way. Comprehensive experiments demonstrate the effectiveness of DropKey for various ViT architectures, *e.g.* T2T and VOLO, as well as for various vision tasks, *e.g.*, image classification, object detection, human-object interaction detection and human body shape recovery.

1 Introduction

Vision Transformer (ViT) [6] has achieved great success for various vision tasks, *e.g.*, image recognition [31; 32; 19; 7; 12], object detection [1], human body shape estimation [17], etc. Prior works mainly focus on researches of patch division, architecture design and task extension. However, the dropout technique for self-attention layer, which plays the essential role to achieve good generalizability, is surprisingly ignored by the community.

Different from the counterpart for Convolutional Neural Networks (CNNs), the dropout in ViT directly utilizes the one in original Transformer designed for Natural Language Processing, which sets attention weights as the manipulation unit with a constant dropout ratio for all layers. Despite of its simplicity, this vanilla design faces three major problems. First, it breaks the probability distribution of attention weights due to the averaging operation on non-dropout units after softmax normalization. Although this regularizes the attention weights, it still overfits specific patterns locally due to the failure on penalizing score peaks, as shown in Fig. 1 (a) and (b); Second, the vanilla design is sensitive to the constant dropout ratio, since high ratio occurs missing of semantic information in

*This work was done during Bonan’s internship at MT Lab, Meitu

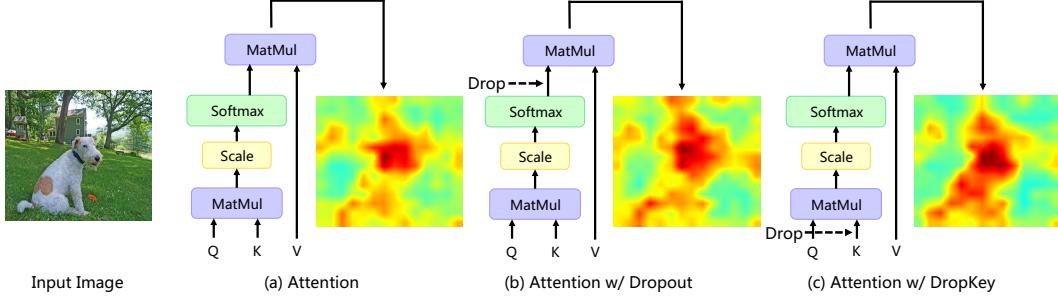


Figure 1: Comparison between the proposed DropKey and existing vanilla dropout techniques in self-attention layers for ViTs. (a) Self-attention without dropout, which suffers overfittings problem to local patch; (b) Self-attention with vanilla dropout, which regularizes the attention weights but still overfits specific patterns; (c) Self-attention with our DropKey, which overcomes prior problems and improves the model to capture vital information in a global manner. Best viewed in color.

high-level representations while low ratios overfitting in low-level features, resulting in the unstable training process; Third, it ignores the structured characteristic of input patch grid to ViT, which plays an effective role to improve performance with blockwise dropout in CNNs. These three problems degrade the performance and limit the generalizability of ViTs.

Motivated by the above, we propose to analyze and improve the dropout technique in self-attention layer, further pushing forward the frontier of ViTs for vision tasks in a general way. In particular, we focus on three core aspects in this paper:

What to drop in self-attention layer Different from dropping attention weights as in the vanilla design, we propose to set the Key as the dropout unit, which is essential input of self-attention layer and significantly affects the output. This moves the dropout operation forward before calculating the attention matrix as shown in Fig. 1 (c) and yields a novel dropout-before-softmax scheme. This scheme regularizes attention weights and keeps their probability distribution at the same time, which intuitively helps penalize weight peaks and lift weight foots. We theoretically verify this property via implicitly introducing an adaptive smoothing coefficient for the attention operator from the perspective of gradient optimization by formulating a Lagrange function. With the dropout-before-softmax scheme, self-attention layers can capture vital information in a global manner, thus overcoming the overfittings problem to specific patterns occurred in the vanilla dropout and enhancing the model generalizability as visualization of feature map in Fig. 1 (c). For the training phase, this scheme can be simply implemented by swapping the operation order of softmax and dropout in vanilla design, which provides a general way to effectively enhance ViTs. For the inference phase, we conduct an additional finetune phase to align the expectations to the training phase, further improving the performance.

How to schedule the drop ratio In contrast to exploiting a constant drop ratio for all layers, we present a new linear decreasing schedule that gradually decreases the drop ratio along the stack of self-attention layers. This schedule leads to a high drop ratio in shallow layers while the low one in deep layers, thus avoiding overfittings to low-level features and preserving sufficient high-level semantics. We experimentally verify the effectiveness of the proposed decreasing schedule for drop ratio to stable the training phase and improve the robustness.

Whether need to perform structured drop Inspired by the DropBlock [10] method for CNNs, we implement two structured versions of the dropout operation for ViTs: the block-version dropout that drops keys corresponding to contiguous patches in images or feature maps; the cross-version dropout that drops keys corresponding to patches in horizontal and vertical stripes. We conduct thorough experiments to validate their efficacy and find that the structure trick useful for CNN is not essential for ViT, due to the powerful capability of ViT to grasp contextual information in full image range.

Given exploration on the above three aspects, we present a novel DropKey method that utilizes Key as the drop unit and decreasing schedule for drop ratio. In particular, DropKey overcomes drawbacks of the vanilla dropout technique for ViTs, improving performance in a general and effective way. Comprehensive experiments on different ViT architectures and vision tasks demonstrate the efficacy of DropKey. Our contributions are in three folds: First, to our best knowledge, we are the first to theoretically and experimentally analyze dropout technique for self-attention layers in ViT from three core aspects: drop unit, drop schedule and structured necessity; Second, according to our analysis, we present a novel DropKey method to effectively improve the dropout technique in ViT. Third, with DropKey, we improve multiple ViT architectures to achieve new SOTAs on various vision tasks.

2 Related Work

Vision Transformers Inspired by the Transformer architectures in NLP, some works introduce attention block to replace convolution layers to model long-range dependencies [6; 34; 27; 29; 28; 11]. ViT[6] is the pioneering work which splits an image to non-overlapping patches and then these patches are fed to a transformer to evaluate attention scores. Prior works mainly focus on patch division, architecture design and tasks extension. For example, CvT[29] proposes a hybrid architecture to mix convolutional with attention layers to introduce local inductive bias. [19; 3; 7] proposes to reduces the computation cost. Despite the progress, previous works ignore the dropout technique in self-attention layer, which play an important role to keep good generalizability.

Dropout Dropout is a common technique for improving the generalizability of neural networks, *e.g.*, CNNs [16], RNNs [20] and GNNs [23]. For CNNs, [10] points out the lack of success of dropout for convolutional layers is due to dense information flowing and then proposes a form of structured Dropout. DropConnect [26] randomly masks a subset of weights within the model. For RNNs, the first prominent research on dropout is presented in [8], which is equipped with a learnable dropout rate.

For GNNs, to alleviate over-fitting and over-smoothing issues, DropEdge [22] randomly removes part of edges in the input stage at each training epoch and theoretically demonstrates the effect of the proposed method. For ViTs, most original publicly code of multiple popular vision transformer simply intuitively apply dropout operation on attention matrix.

DropAttention [33] presents a novel way for transformer in NLPs. It refers to perform randomly drop on the attention matrix. There are two main forms of applying DropAttention for the transformer training, DropAttention-unit and DropAttention-elements. Similar to the standard Dropout [24], DropAttention-unit randomly drops the unit, that is, these dropped token will not be used as attend patches for each query patch. As the general form of DropAttention-unit, DropAttention-elements randomly drops elements in attention weights matrix which is similar to DropConnect [26]. Different from standard Dropout [24], DropAttention introduce re-normalize to guarantee the sum of attention weights remain 1 and help training process to be more steady. However, there is no further discussion and theoretical analysis on this phenomenon.

3 Method

Our generic DropKey is inspired by DropAttention and proposed for self-attention operator in vision transformer. The main idea of DropKey is to adaptively adjust attention weight to obtain a smoother attention vector. In this section, we start by introducing the theoretical explanation of DropKey and then the implementation will be explained in detail.

3.1 Methodology

As discussed in the above section, transformer-based model tends to rely on local features rather than general global information. To alleviate this issue, we propose to reduce local-bias by encouraging models to learn a smoother attention weight for each patch. To this end, we attempt to reduce the attention weight of the patch which has a large attention weight, and vice versa. Nevertheless, it is tedious and difficult to achieve this by explicitly setting rules. In this work, we found DropKey achieves the above implicitly by normalizing the attention vector that has performed the dropout

operation. Specifically, by given an image $I \in \mathbb{R}^{H \times W \times C}$, where H denotes height, W denotes width and C denotes channels, Vision Transformer architecture starts by dicing it into $n_h \times n_w$ patches $x \in \mathbb{R}^{n_c}$. Then, x is used as input of self- attention layer and the output o can be computed as follows

$$o = \sum_{j=1}^{n_h n_w} \left(\frac{d_j p_j}{\sum_{j=1}^{n_h n_w} d_j p_j} \right) v_j \quad (1)$$

$$p_j = \frac{\exp(\frac{q k_j^T}{scale})}{\sum_{j=1}^{n_h n_w} \exp(\frac{q k_j^T}{scale})} \quad (2)$$

where $q = \mathcal{F}_q(x)$, $k = \mathcal{F}_k(x)$, $v = \mathcal{F}_v(x)$, and q_i , k_j , v_j denote query of i^{th} patch, key of j^{th} patch, value of j^{th} patch and o denotes the output of one patch. $\mathcal{F}_q(\cdot)$, $\mathcal{F}_k(\cdot)$, $\mathcal{F}_v(\cdot)$ denote projection layers with weights of dimensions $n_c \times n_c$. $scale$ denotes scaling factor and is set to $\sqrt{n_c}$. d denotes drop ratio and $d_j \sim \text{Bernoulli}(1 - d)$. Here, for convenience, we focus on one head and omit the index of patch. By performing re-normalize, DropKey has the ability to adaptively adjust attention weight to smoother. Next, we provide theoretical analysis of our method to demonstrate its effectiveness.

Here, we formulate the expectation output of model via introducing DropKey in training stage as:

$$\begin{aligned} \mathbb{E}_{d_j, 1 \leq j \leq n_h n_w}[o] &= \mathbb{E}_{d_j, 1 \leq j \leq n_h n_w} \left[\sum_{j=1}^{n_h n_w} \left(\frac{d_j p_j}{\sum_{j=1}^{n_h n_w} d_j p_j} \right) v_j \right] \\ &= \sum_{j=1}^{n_h n_w} \mathbb{E}_{d_j, 1 \leq j \leq n_h n_w} \left[\frac{d_j}{\sum_{j=1}^{n_h n_w} d_j p_j} \right] p_j v_j = \sum_{j=1}^{n_h n_w} c_j p_j v_j \end{aligned} \quad (3)$$

where $c_j = \mathbb{E}_{d_j, 1 \leq j \leq n_h n_w} \left[\frac{d_j}{\sum_{j=1}^{n_h n_w} d_j p_j} \right] > 0$ is a smoothing coefficient which is related to d_j and p_j .

Note that c in Equ 3 can be considered to add an additional smoothing prior to distribution p , i.e., $c_s < c_t$ when $p_s > p_t$, $1 \leq s, t \leq n_h n_w$. The specific proof is as follows:

$$\begin{aligned} c_s - c_t &= \mathbb{E}_{d_j, 1 \leq j \leq n_h n_w} \left[\frac{d_s}{\sum_{j=1}^{n_h n_w} d_j p_j} \right] - \mathbb{E}_{d_j, 1 \leq j \leq n_h n_w} \left[\frac{d_t}{\sum_{j=1}^{n_h n_w} d_j p_j} \right] \\ &= d(1 - d) \mathbb{E}_{d_j, 1 \leq j \leq n_h n_w} \left[\frac{-1}{\sum_{j \neq s, t}^{n_h n_w} d_j p_j + p_t} \right] + d(1 - d) \mathbb{E}_{d_j, 1 \leq j \leq n_h n_w} \left[\frac{1}{\sum_{j \neq s, t}^{n_h n_w} d_j p_j + p_s} \right] \\ &= d(1 - d) \mathbb{E}_{d_j, 1 \leq j \leq n_h n_w} \left[\frac{1}{\sum_{j \neq s, t}^{n_h n_w} d_j p_j + p_s} - \frac{1}{\sum_{j \neq s, t}^{n_h n_w} d_j p_j + p_t} \right] < 0 \end{aligned} \quad (4)$$

It is natural to find c serves as a factor to implicitly encourage the model properly to reduce the consideration of the patch with large p and improve the effectiveness of the patch with small attention weight. Meanwhile, c can be adjusted adaptively according to the distribution of samples in the training stage without any manual design. Nevertheless, a noteworthy problem is that since we removed DropKey in the inference phase, the output expectations in training and inference stage are inconsistent, which will decrease the performance. In Section 4.1.1, we propose two methods to alleviate this problem.

Having explored the implicit regularization effects of DropKey, we also demonstrate it from the perspective of gradient optimization. For simplicity, we start by considering a simple but universal optimization objective as follows:

$$\min_{p_j, v_j} \frac{1}{2} \left\| \sum_{j=1}^N p_j v_j - y \right\|^2 \quad \text{s.t. } p_j > 0, \sum_{j=1}^N p_j = 1, 1 \leq j \leq N \quad (5)$$

where $p_j \in \mathbb{R}$, $v_j \in \mathbb{R}^r$ denotes learnable parameters and $y \in \mathbb{R}^r$ denotes target. Here, p_j and v_j can be considered as attention weight and value in attention mechanism. Naturally, v_j can be decomposed into two directions, which are the same direction as y and the direction perpendicular to y :

$$v_j = \beta_j e + \alpha_j \quad (6)$$

where β is a scalar, $e \in \mathbb{R}^r$ is the unit vector of y and α_j is the component perpendicular to y . Similarly, we can rewrite y to Me where $M = ||y||$. Then, the Lagrange function of Equ 5 can be formulated as follows:

$$L(p_j, v_j, \lambda) = \frac{1}{2} \left\| \sum_{j=1}^N p_j v_j - y \right\|^2 - \lambda \left(\sum_{j=1}^N p_j - 1 \right) \quad (7)$$

In order to analysis the gradient, we take the partial derivatives of p_j and v_j :

$$\frac{\partial L}{\partial p_j} = (v - y)^T v_j - \lambda = \left(\sum_{j=1}^N p_j \beta_j - M \right) \beta_j + \left(\sum_{j=1}^N p_j \alpha_j \right)^T \alpha_j - \lambda \quad (8)$$

$$\frac{\partial L}{\partial v_j} = p_j (v - y) = p_j \left(\sum_{j=1}^N p_j \beta_j - M \right) e + p_j \left(\sum_{j=1}^N p_j \alpha_j \right) \quad (9)$$

Here, for simplicity, we note $v = \sum_{j=1}^N p_j v_j = (\sum_{j=1}^N p_j \beta_j)e + \sum_{j=1}^N p_j \alpha_j$. From the chain rule of backpropagation, we also have:

$$\frac{\partial L}{\partial \beta_j} = p_j \left(\sum_{j=1}^N p_j \beta_j - M \right) \quad \frac{\partial L}{\partial \alpha_j} = p_j \left(\sum_{j=1}^N p_j \alpha_j \right) \quad (10)$$

Equ 10 indicates that the gradient $\frac{\partial L}{\partial \beta_s} < \frac{\partial L}{\partial \beta_t}$ when $p_s > p_t$ and this backward propagation properties would lead $\frac{\partial L}{\partial p_s} < \frac{\partial L}{\partial p_t}$. By subtracting $\frac{\partial L}{\partial p_s}$ and $\frac{\partial L}{\partial p_t}$, we will have:

$$\frac{\partial L}{\partial p_s} - \frac{\partial L}{\partial p_t} = \left(\sum_{j=1}^N p_j \beta_j - M \right) (\beta_s - \beta_t) + \left(\sum_{j=1}^N p_j \alpha_j \right)^T (\alpha_s - \alpha_t) \quad (11)$$

Since all parameters are randomly initialized, it can be assumed that the initial solution is far from the optimal solution, that is, $p_1^{(0)} = p_2^{(0)} = \dots = p_N^{(0)} = \frac{1}{N}$, $|\beta_j^{(0)}| \ll M$, $1 \leq j \leq N$. Consequently, the first additive term plays the leading role in Equ 11. Due to $\frac{\partial L}{\partial \beta_s} < \frac{\partial L}{\partial \beta_t}$, the update speed of β_s is faster than β_t and then lead $\beta_s > \beta_t$, $\frac{\partial L}{\partial p_s} < \frac{\partial L}{\partial p_t}$. Based on the above analysis, it can be concluded that a larger p would promote β larger, while a larger β would further promote p larger. Finally, the output of model would be controlled by a few sparse blocks. Conversely, DropKey avoids suffering local-bias by introducing the parameter c to enforce distribution p to be smoother.

3.2 Implementation

What to drop? Different from DropAttention, we integrate Dropout and re-normalize into one stage by dropping key rather than weight. At each training iteration, DropKey masks a certain rate of keys of the input key map by random. It is worth noting that we generate masked key map for each query, instead of sharing the same masked key map for all query vectors. Specifically, given the query $Q \in \mathbb{R}^{n_h n_w \times n_c}$, key $K \in \mathbb{R}^{n_h n_w \times n_c}$ and value $V \in \mathbb{R}^{n_h n_w \times n_c}$ of a feature map, it first computes the dot products of the query with all keys and divide each by scaling factor $scale = \sqrt{n_c}$. Subsequently, we randomly generate a mask matrix $D \in \mathbb{R}^{1 \times (n_h n_w \times n_h n_w)}$ with drop ratio d to enforce some elements of the similarity matrix to be -inf. The formulation of D as follows:

$$d_j = \begin{cases} 0 & \text{with probability } 1 - d \\ -\infty & \text{with probability } d \end{cases} \quad (12)$$

Finally, the attention weight matrix is computed by given masked similarity matrix as input. Formally, we compute the outputs of a patch as:

$$o = \sum_{j=1}^{n_h n_w} p_j v_j \quad (13)$$

Algorithm 1 Attention with DropKey code

```

1   # N: token number, D: token dim
2   # Q: query (N, D), K: key (N, D), V: value (N, D)
3   # use_DropKey: whether use DropKey
4   # mask_ratio: ratio to mask
5
6   def Attention(Q, K, V, use_DropKey, mask_ratio)
7       attn = (Q * (Q.shape[1] ** -0.5)) @ K.transpose(-2, -1)
8
9       # use DropKey as regularizer
10      if use_DropKey == True:
11          m_r = torch.ones_like(attn) * mask_ratio
12          attn = attn + torch.bernoulli(m_r) * -1e12
13
14      attn = attn.softmax(dim=-1)
15      x = attn @ V
16      return x

```

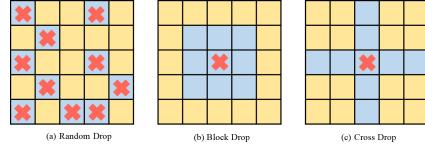


Figure 2: Mask sampling in DropKey, DropKey-Block and DropKey-Cross. The yellow patches are used as attend key to interact with a query and blue patches are dropped. Red symbol denotes the valid seed and the window size of DropKey-Block and DropKey-Cross is 3 and 1, respectively.

$$p_j = \frac{\exp(d_j + \frac{qk_j^T}{scale})}{\sum_{j=1}^{n_h n_w} \exp(d_j + \frac{qk_j^T}{scale})} \quad (14)$$

where q_i , k_j , v_j denote query of i^{th} patch, key of j^{th} patch, value of j^{th} patch. It is not difficult to find that Equ 13 and Equ 14 is the same equivalent form of Equ 1 and Equ 2. In particular, as shown in Algorithm 1, we only need to add two lines of code to pure attention to implement DropKey. Obviously, compared with Dropout, DropKey only changes the position of introducing mask.

Scheduled Drop Ratio Vision transformer always consists of few self-attention blocks to gradually learn high-dimensional features. Generally, early layers operates low-level visual features and deeper layers aim to model spatially coarse but complex information. Hence, we attempt to set smaller drop ratio for deeper layers to avoid missing important objects-relevant information. Specifically, instead of stochastically dropping with some fixed probability in Dropout, we gradually decrease number of dropped keys over the layers during training stage. We find that this scheduled drop ratio not only work well for DropKey, but also significantly improves the performance of Dropout.

Structured DropKey Although structured drop has been detailedly explored in convolution network, there is no work to study the impact of structured drop on vision transformer. In this section, we implement two structured forms of DropKey which are named DropKey-Block and DropKey-Cross (see Figure 2). For DropKey-Block, inspired by DropBlock [10], we drop contiguous patches in a square-shape window of feature maps. Actually, we re-set drop ratio d_{block} for each patch as: $d_{block} = \frac{d}{s^2} \frac{n_h n_w}{(n_h - s + 1)(n_w - s + 1)}$ where d denotes the probability of dropping a patch in DropKey. The valid seed region is $(n_h - s + 1)(n_w - s + 1)$ where s denotes the size of window and n_h, n_w denotes height, weight of feature map.

The cross-shape window attention is proved to achieve strong modeling capability [5] which also indicates the information in this structure is correlated. For DropKey-Cross, we discard features in a cross window to prevent information flow through self-attention. In our implementation, we drop the rows and columns of valid seed and re-set the drop ratio d_{cross} for each patch as: $d_{cross} = \frac{d}{s(n_h + n_w - s)} \frac{n_h n_w}{(n_h - s + 1)(n_w - s + 1)}$. Note that the formulations for block- and cross-version are only the approximation, because there will be some overlapped when perform drop.

Align Expectation As mentioned in Section 3.1, the misaligned expectations have a certain negative impact on the model, so we attempt to use two methods to align the expectation. The first one is to use Monte Carlo method to estimate c . We perform multiple random drop and calculate the attention weight matrix after each drop operation. Finally, the average of calculated multiple weight matrices is applied as the input for the next step. For the second one, we take the inspiration from [9] and propose to finetune the model without DropKey, as an extra stage after DropKey training. We experimentally verify that the second strategy performs better.

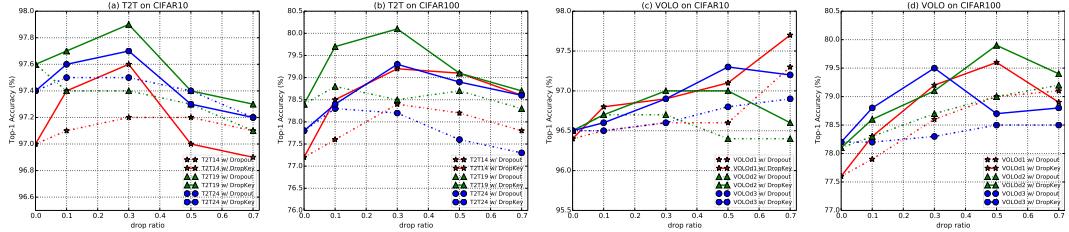


Figure 3: Comparison of models with Dropout or DropKey on CIFAR10 and CIFAR100.

4 Experiments

We conduct experiments on three tasks, image classification, object detection, human-object interaction detection and human mesh recovery, to show the efficacy and generalizability of our DropKey for improving ViTs.

4.1 Image Classification

Datasets We conduct the following experiments with T2T [31] and VOLO [32] for image classification on CIFAR10 [13], CIFAR100 [15] and ImageNet [4]. a) We conduct the ablation study to demonstrate the effects of introducing modified DropAttention to three T2T and three VOLO backbone architectures on CIFAR10 and CIFAR100. b) We validate the modified DropAttention by training T2T and VOLO from scratch on ImageNet.

Training By default, models are trained with random initialization on all datasets. Due to the missing training setting for T2T [31] and VOLO [32], we conduct grid search on learning rate and epoch number to make the vanilla backbones achieve the best performance. All implementation specifics can be seen in Appendix. For ImageNet, our training recipe follows [31; 32]. Specifically, we set 50 epochs and keep learning rate as 10^{-5} for finetune stage for all datasets.

4.1.1 Ablations on CIFAR10 and CIFAR100

Does DropKey boost generalization? T2T [31] and VOLO [32] are the widely used vision transformer architecture for image recognition. However, the DropAttention applied in these architectures does not apply re-normalize operation. In the following experiments, we plug in the DropKey to the family of T2T/VOLO and compare results with the version that use Dropout. The results in Figure 3 indicate DropKey can consistently increase the performance of various architectures on CIFAR10 and CIFAR100. On CIFAR100 with T2T19, for example, DropKey gains 1.7% improvement for the adaptive coefficient yields further improvement. Additionally, we found that drop ratio is also an essential factor affecting performance. Take T2T14 with DropKey as an example, the accuracy initially increases with increasing drop ratio until it achieves its peak accuracy of 79.2% at drop ratio = 0.3 and it declines upon further increase of the drop ratio. These suggest that the introduction of DropKey would be beneficial to the network for avoiding suffering overfits. However, a larger drop ratio can lead to an inability to capture valid information about the object. Similar to DropKey, the performance of Dropout is also related to drop ratio.

How does the choice of scheduled strategy of drop ratio impact accuracy? The ablation in Table 1 analyzes the accuracy of DropKey by adjusting the strategy of setting drop ratio. Specifically, we validate three strategy as follows: a) Constant: drop ratio is constant over self attention layers. b) Scheduled \uparrow : drop ratio is linearly increased over self attention layers. c) Scheduled \downarrow : drop ratio is linearly decreased over self attention layers. Firstly, we find that Scheduled \downarrow outperform Constant whether using re-normalize or not. Secondly, introducing drop ratio with Scheduled \uparrow would seriously affect the performance of the model and even perform worse than the pure vision transformer. This phenomenon is in line with our expectation, since deeper layers often contain high-level semantic information which is essential to perform classification. A large drop ratio in deep layer will increase the risk of losing important features which would make model difficult to converge to a generalized solution. Hence, we set drop ratio with Scheduled \downarrow in all following experiments.

Table 1: Comparison of models with Dropout or DropKey on CIFAR10 and CIFAR100 when introducing different scheduled strategy of drop ratio. The drop ratio is set as 0.3 for all models.

Model	CIFAR10			CIFAR100		
	Scheduled↑	Constant	Scheduled↓	Scheduled↑	Constant	Scheduled↓
T2T14 [31] + Dropout	96.7	96.9	97.2	77.0	77.8	78.4
T2T14 [31] + DropKey	97.0	97.4	97.6	77.2	78.6	79.2
T2T19 [31] + Dropout	96.6	97.1	97.4	77.3	78.1	78.5
T2T19 [31] + DropKey	97.1	97.2	97.9	78.0	78.4	80.1
T2T24 [31] + Dropout	96.4	97.0	97.5	76.7	77.3	78.2
T2T24 [31] + DropKey	96.6	97.2	97.7	76.9	78.3	79.3
VOLOd1 [32] + Dropout	95.8	96.4	96.6	77.3	78.2	79.0
VOLOd1 [32] + DropKey	96.2	96.7	97.1	78.1	78.6	79.6
VOLOd2 [32] + Dropout	95.6	96.1	96.4	77.0	78.4	79.0
VOLOd2 [32] + DropKey	95.7	96.6	97.0	77.7	79.0	79.9
VOLOd3 [32] + Dropout	96.0	96.3	96.6	77.1	77.9	78.5
VOLOd3 [32] + DropKey	96.2	96.4	96.9	77.4	78.1	78.7

Table 2: Comparison of different structured drop on CIFAR100. s denotes window size and drop ratio is set as 0.3 for all models.

Model	Random	Block		Cross	
		$s=3$	$s=5$	$s=1$	$s=3$
T2T14 [31] + Dropout	78.4	78.1	77.1	77.9	76.8
T2T14 [31] + DropKey	79.2	78.7	77.5	78.6	77.4
VOLOd1 [32] + Dropout	78.6	78.0	76.7	77.6	76.4
VOLOd1 [32] + DropKey	79.2	78.5	77.5	78.3	77.1

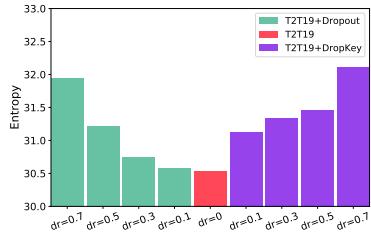


Figure 4: The histogram of entropy on CIFAR100. dr denotes drop ratio.

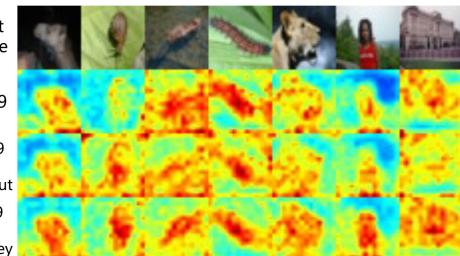


Figure 5: Visualization on CIFAR100.

Is Dropkey helpful to capture global information? The contribution of DropKey is to generate a smoother attention weight matrix to encourage models to focus on global features. Class token has been proved to be effective in aggregating the information of the whole image, so we measure the smoothness of the attention distribution by calculating the entropy of its attention weight vector. Specifically, we define the entropy of the attention vector of class token as follows $E_i = - \sum_j p_j^i \log p_j^i$ where p_j^i denotes the attention weight of j^{th} attending patch in i^{th} head.

Note that a small entropy denotes that this head focuses on sparse patches. For convenience, the mean of entropy of multi-head as E is reported in Figure 4. Clearly, it reads that while both Dropout and DropKey are able to improve the value of entropy, the improvement by DropKey is more significant. Meanwhile, when the drop rate increases, the entropy increase accordingly. In addition, to further verify the necessity of the adaptive coefficient for capturing general global information, we visualize the attention map (Figure 5) from class token in the last transformer layer. Obviously, the model with DropKey smoothly assigns attention weight to the area related to the instance.

Is Dropkey help model robust to occlusions? To further verify that DropKey can alleviate local-bias problems, we study whether vision transformer with DropKey perform robustly in occluded scenarios, where some content of the image is missing. Specifically, a subset of patches is randomly selected and dropped before input to self-attention layer. In this experiment, with the drop ratio found in the manuscript, we randomly drop 10%, 30%, 50%, 70% and 90% of the patches to test the performance of trained models. For convenience, we define information loss as the ratio of

Table 3: Robustness against occlusion in images is studied under T2T19 and T2T24 on CIFAR100. The drop ratio is set as 0.3 for all models.

Information Loss	0.0	0.1	0.3	0.5	0.7	0.9
T2T19 [31]	78.4	77.6 ^{-0.8}	75.6 ^{-2.8}	73.1 ^{-5.3}	65.1 ^{-13.3}	28.2 ^{-50.2}
T2T19 [31] + Dropout	78.5	77.4 ^{-1.1}	76.5 ^{-2.0}	73.9 ^{-4.6}	66.0 ^{-12.5}	30.2 ^{-48.3}
T2T19 [31] + DropKey	80.1	79.4 ^{0.7}	78.2 ^{-1.9}	75.8 ^{-4.3}	68.7 ^{-11.4}	32.7 ^{-47.4}
T2T24 [31]	77.8	77.1 ^{-0.7}	75.2 ^{-2.6}	72.7 ^{-5.1}	63.5 ^{-14.3}	23.7 ^{-54.1}
T2T24 [31] + Dropout	78.2	77.3 ^{-0.9}	75.7 ^{-2.5}	72.8 ^{-5.4}	65.3 ^{-12.9}	27.9 ^{-50.3}
T2T24 [31] + DropKey	79.3	78.6 ^{-0.7}	78.0 ^{-1.3}	74.3 ^{-5.0}	66.9 ^{-12.4}	29.7 ^{-49.6}

Table 4: Comparison of models w/ or w/o Align Expectations on CIFAR100. Pure denotes the model without align expectation and drop ratio is set as 0.3 for all models.

Model	Pure	w/ Monte Carlo		w/ Finetune
		2000	6000	
T2T14 [31] + Dropout	78.4	-	-	78.4
T2T14 [31] + DropKey	78.8	78.9	79.1	79.2
T2T19 [31] + Dropout	78.5	-	-	78.6
T2T19 [31] + DropKey	79.3	79.5	80.2	80.1

dropped and total patches. To eliminate the influence of random occlusions, we report the mean of accuracy across 5 runs. The results reported in Table 3 show significantly robust performance of the model with DropKey against its with Dropout trick. For example, T2T24+Dropout achieves 75.7% accuracy in comparison to T2T24+DropKey which obtains 78.0% accuracy when 30% of the patches are removed. A surprising phenomenon can be observed that when 90% of the image information is randomly dropped, T2T19+DropKey and T2T24+DropKey still exhibits 32.7% and 29.7% accuracy, respectively. Consequently, compared with Dropout, models with DropKey show significant robustness to the content removal.

Is structured drop is useful to vision transformer? In this section, we analyze whether structured masking is useful for vision transformer via DropKey-Block and DropKey-Cross. Specifically, we swept over window size from 3 to 5 for DropKey-Block and from 1 to 3 for DropKey-Cross. Compared with the DropKey, the structured drop degrade the performance of the model in all case and classification accuracy will decrease with the further increase of the window size. One possible reason for this phenomenon is that a larger drop ratio in shallow transformer layer results in losing vital information of object. However, a larger drop ratio is the key to avoid overfit to low-level feature. Therefore, we encourage set a larger drop ratio rather than the introduction of structured drop.

How does align expectations impact accuracy? As discussed above, the mismatched expectations problem arises when directly use trained model with DropKey to test. Herein, to address above issue, we explore the impact of two methods, Monte Carlo and finetune, on the performance of network. The results in Table 4 indicates that accuracy can be improved regardless of the expectation alignment operation. Additionally, the accuracy of Monte Carlo increases with increasing sampling iterations. However, huge sampling iterations will lead to unacceptable computational cost. To fairly compare with Dropout, we also finetune it with the same training hyper-parameters as DropKey and then note that finetune only brings insignificant performance improvement or even over-fitting. In conclusion, alignment expectation of training and inference stage indeed can further improve the accuracy of the model.

4.1.2 DropKey vs Dropout on ImageNet

We provide the accuracy under other different drop ratio for each backbone with DropKey on ImageNet, and contrast them with existing State of the Arts (SOTA) equipped with Dropout, including T2T14, T2T19, VOLOd1 and VOLOd2 in Table 5. We have these findings: (1) Clearly, our DropKey obtains significant enhancement against pure backbone. Meanwhile, We note that Dropout can only

Table 5: Comparisons on backbones with Dropout or DropKey on ImageNet. dr denotes drop ratio.

Model	Top-1	Model	Top-1
T2T14 [31]	81.76 ± 0.05	VOLOd1 [32]	84.27 ± 0.12
T2T14 [31] + Dropout(dr=0.05)	81.65 ± 0.04	VOLOd1 [32] + Dropout(dr=0.05)	84.35 ± 0.11
T2T14 [31] + Dropout(dr=0.1)	81.53 ± 0.08	VOLOd1 [32] + Dropout(dr=0.1)	84.31 ± 0.04
T2T14 [31] + DropKey(dr=0.05)	82.04 ± 0.14	VOLOd1 [32] + DropKey(dr=0.05)	84.53 ± 0.03
T2T14 [31] + DropKey(dr=0.1)	81.93 ± 0.09	VOLOd1 [32] + DropKey(dr=0.1)	84.39 ± 0.02
T2T19 [31]	82.56 ± 0.09	VOLOd2 [32]	85.24 ± 0.05
T2T19[31] + Dropout(dr=0.05)	82.63 ± 0.05	VOLOd2 [32] + Dropout(dr=0.05)	85.22 ± 0.08
T2T19[31] + Dropout(dr=0.1)	82.68 ± 0.04	VOLOd2 [32] + Dropout(dr=0.1)	85.21 ± 0.03
T2T19[31] + DropKey(dr=0.05)	82.71 ± 0.02	VOLOd2 [32] + DropKey(dr=0.05)	85.33 ± 0.12
T2T19[31] + DropKey(dr=0.1)	82.94 ± 0.04	VOLOd2 [32] + DropKey(dr=0.1)	85.38 ± 0.06

Table 6: Comparison of DETR with Dropout and DropKey on COCO validation set.

Table 7: Comparison of METRO with Dropout and DropKey on HUMBI.

Model	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
DETR [1]	41.9	62.1	44.2	20.6	45.7	60.7
DETR [1] + Dropout	42.2	62.3	44.3	20.8	45.9	61.1
DETR [1] + DropKey	42.9	63.4	44.7	21.1	46.7	61.8

Model	mPVE _↓	mPJPE _↓	PA-mPJPE _↓
METRO [17]	57.8	51.5	39.9
METRO [17] + Dropout	57.9	51.6	39.7
METRO [17] + DropKey	57.1	51.0	39.5

bring a little boost to backbones and it even would degrade the performance of the model in some cases (eg. T2T14). This indicates that the adaptive smoothing operator is more helpful to encourage the model to capture general information in large-scale datasets. (2) Compare to lightweight models, heavyweight often need the larger drop ratio to achieve a remarkable boost. For T2T14 and VOLOd1, the best accuracy is obtained under the dr=0.05. Instead, T2T19 and VOLOd2 achieve the best accuracy when dr=0.1, which suggests heavyweight models adopt a larger drop ratio to prevent overfitting.

4.2 Object Detection in COCO

In this section, we verify the generalizability and effectiveness of DropKey for object detection on COCO dataset [18]. We apply DETR [1] framework for the experiments. Specifically, DETR is composed of the backbone, encoder and decoder. We followed the model architecture, anchor definition and training recipe in [1] to build DETR+Dropout and DETR+DropKey. In Table 6, we report the results of DETR [1] with Dropout or DropKey in terms of AP. It can be seen that, our DropKey significantly outperforms existing Dropout. In detail, DropKey achieve +0.7, +1.1, +0.4, +0.3, +0.8 and +0.7 higher AP than the Dropout, which is regarded as a remarkable boost considering the challenge on this benchmark. Another observation is that large objects benefit more from DropKey than small objects. These results suggest that DropKey has the advantage over Dropout.

4.3 Human-Object Interaction Detection in HICO-DET

Since the task of scene graph and relation understanding is quite sensitive to global context learning, so we also verify the effectiveness of DropKey in QPIC [25] for Human-Object Interaction Detection in HICO-DET [2]. QPIC is a transformer-based feature extractor that can effectively aggregate contextually important information. We respectively introduce Dropout and DropKey to both encoder and decoder in QPIC and train these models with the same training recipe as [25]. The model named "QPIC-ResNet101 + Dropout (dr=0.1)" is the result of our reproduction of QPIC with *Scheduled* ↓ drop ratio (Official realization is equipped with *Constant* drop ratio). The experimental results in Table 8 suggest DropKey has higher performance gain on the different evaluating indicators which are consistent with our expectations.

4.4 Human Body Mesh Recovery in HUMBI

We verify the efficacy of DropKey for human body mesh task on HUMBI [30]. We follow conventions to train with 294 subjects from scratch and test with 120 subjects. We use Mean Per Vertex Error (MPVVE) [21], Mean Per Joint Position Error (MPJPE) [14] and Procrustes Analysis MPJPE (PA-MPJPE) [35] as metrics. We use public METRO [17] as the baseline to apply DropKey to self-

Table 8: Comparison of QPIC with Dropout and DropKey on HICO-DET.

Model	full (Default)	rare (Default)	non-rare (Default)	full (Known object)	rare (Known object)	non-rare (Known object)
QPIC-ResNet101 [25] + Dropout (dr=0.1)	29.96	24.03	31.63	32.42	26.01	34.31
QPIC-ResNet101 [25] + Dropout (dr=0.3)	30.02	24.17	31.72	32.49	26.13	34.41
QPIC-ResNet101 [25] + DropKey (dr=0.3)	30.87	24.63	32.24	33.21	26.66	34.94

attention layer. Table 7 summaries the results. we can see that DropKey consistently improves the performance for all standard metrics. Specifically, DropKey brings about 0.7, 0.5 and 0.4 on mPVE, mPJPE and PA-mPJPE, respectively, while Dropout only boosts 0.2 performance on one metric. These results further verify the generalizability of DropKey as well as its superiority over vanilla dropout.

We also present visualization on HUMBI in Figure 6 and results show that DropKey can encourage the model to capture dense interactions with highly correlated vertices to learn the robust representation. METRO w/ Dropout only focuses on vertices around elbow joint which ignores global context. For METRO w/ DropKey, it considers the interactions with vertices (eg. wrist joint and arm) which are helpful to predict the precise location of target joints. This further demonstrates that the proposed DropKey can stimulate the model to capture vital information in a global manner.

5 Conclusion

In this paper, we explore the drop unit, drop schedule and structured necessity of the dropout technique in ViT. Specifically, we propose to set Key as the drop unit, which yields a novel dropout-before-softmax scheme. We theoretically and experimentally verify that this scheme can regularize attention weights and meanwhile and keep them as a probability distribution simultaneously, helping capture vital patterns in a global manner and overcome local-bias problems that occurred to vanilla dropout. In addition, we present a new decreasing schedule for drop ratio, which stabilizes the training phase by avoiding overfittings in low-level features and maintaining sufficient high-level features. Moreover, we also experimentally show that structured dropout is not necessary for ViT. We distill the above analysis as a novel DropKey method, which plays as an improved version of dropout for ViT. Comprehensive experiments with different architectures on various vision tasks demonstrate the effectiveness of the proposed DropKey for enhancing ViTs.

References

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.
- [2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, pages 381–389, 2018.
- [3] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS*, 2021.
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [5] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021.
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

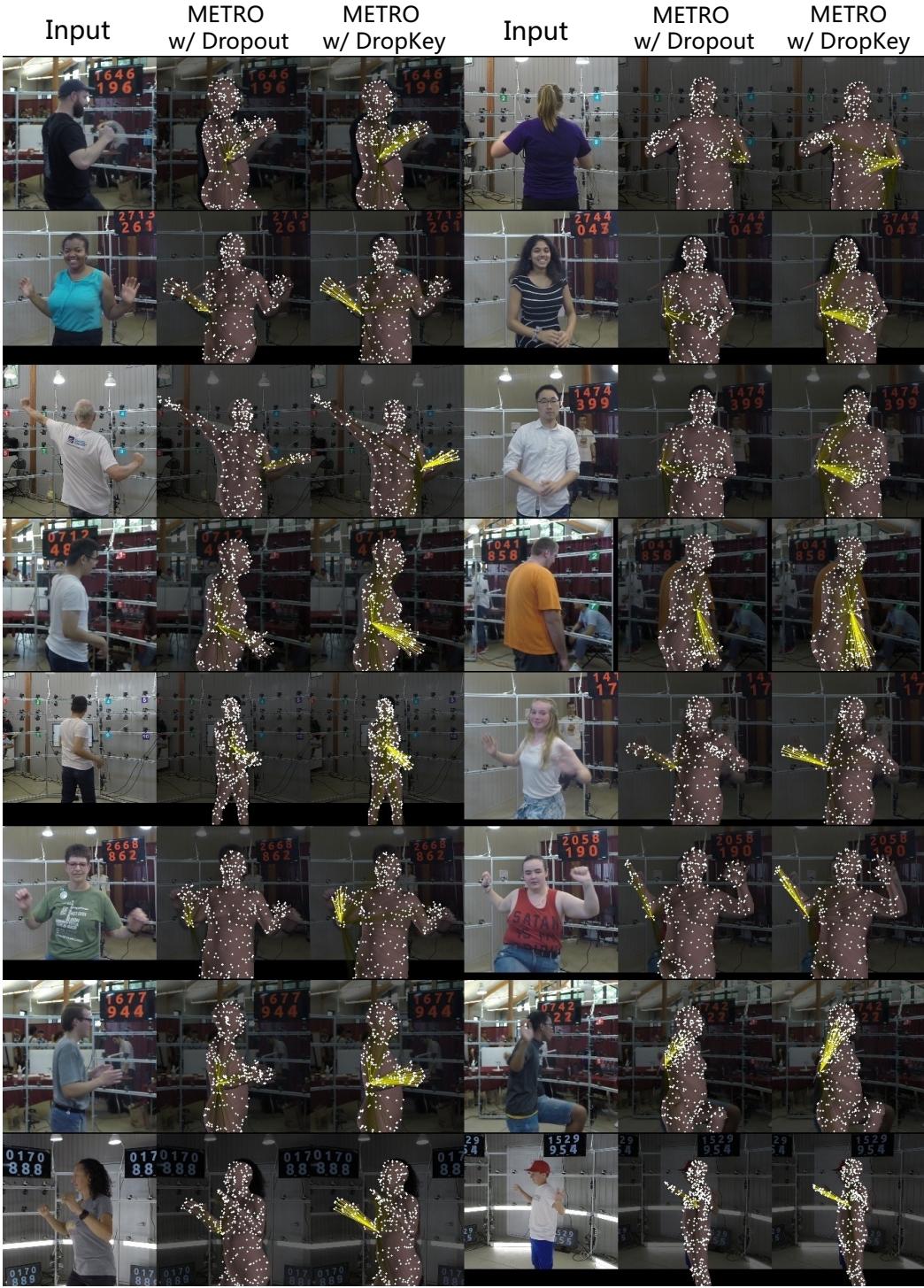


Figure 6: DropKey encourages model to learn robust interactions among body joints and mesh vertices for human mesh reconstruction. Given an input image, METRO w/ Dropout predicts elbow joint only by taking sparse interactions with mesh vertices which is related to elbow into consideration. DropKey encourages model to capture dense interactions with highly correlated vertices to learn the robust representation.

- [7] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021.
- [8] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks, 2016.
- [9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.
- [10] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. In *NeurIPS*, 2018.
- [11] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021.
- [12] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *NeurIPS*, 2021.
- [13] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
- [14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.
- [15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [17] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021.
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *arXiv: Computer Vision and Pattern Recognition*, 2014.
- [19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.
- [20] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.
- [21] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018.
- [22] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.
- [23] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.
- [24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [25] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, pages 10410–10419, 2021.

- [26] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *ICML*, pages 1058–1066, 2013.
- [27] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 568–578, 2021.
- [28] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, pages 8741–8750, 2021.
- [29] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cv_t: Introducing convolutions to vision transformers. In *ICCV*, pages 22–31, 2021.
- [30] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. H umbi: A large multiview dataset of human body expressions. In *CVPR*, pages 2990–3000, 2020.
- [31] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, 2021.
- [32] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *arXiv preprint arXiv:2106.13112*, 2021.
- [33] Lin Zehui, Pengfei Liu, Luyao Huang, Junkun Chen, Xipeng Qiu, and Xuanjing Huang. Dropattention: a regularization method for fully-connected self-attention networks. *arXiv preprint arXiv:1907.11065*, 2019.
- [34] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021.
- [35] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):901–914, 2018.