MEGVII 旷视

Joint Token Pruning and Squeezing Towards More Aggressive Compression of Vision Transformers

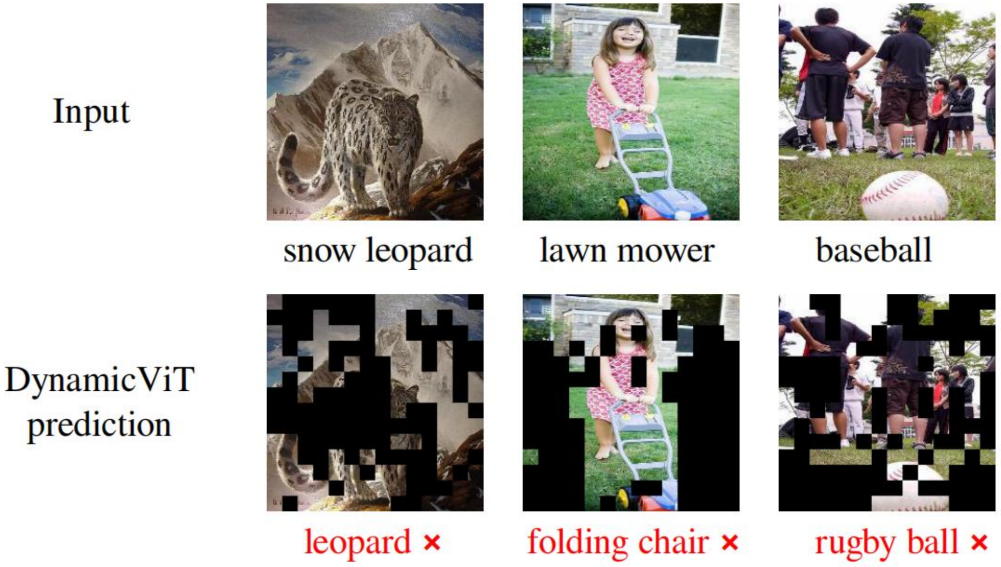
Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, Jiajun Liang

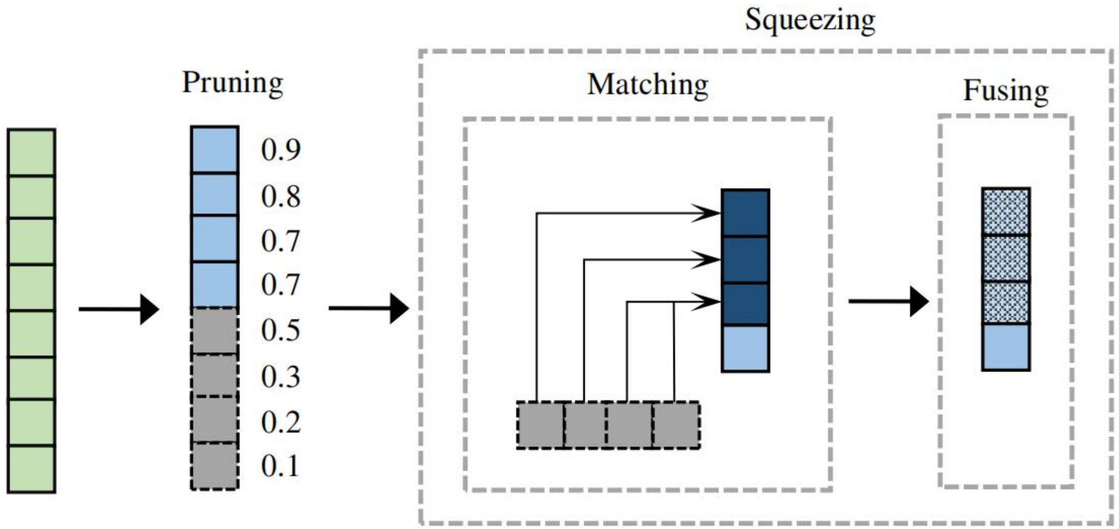
Paper tag: TUE-AM-200

Quick Preview

- For token pruning in vision transformers, discarding tokens leads to incomplete subject and background context loss.

- We propose TPS: a nearest-neighbor matching algorithm to dispatch each pruned token to the most similar reserved token.

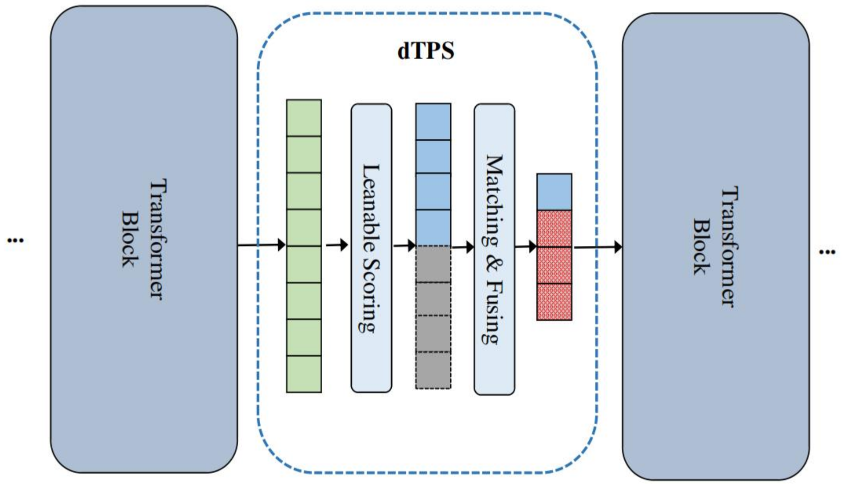


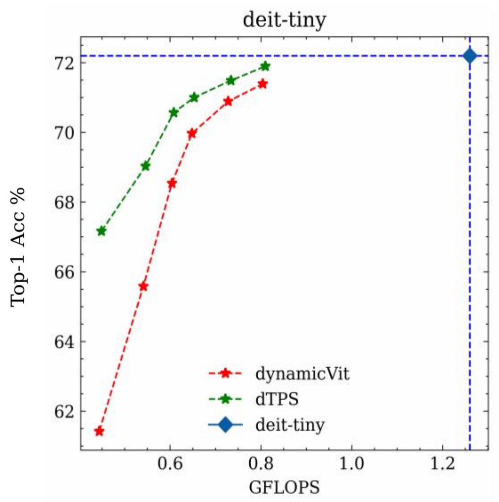


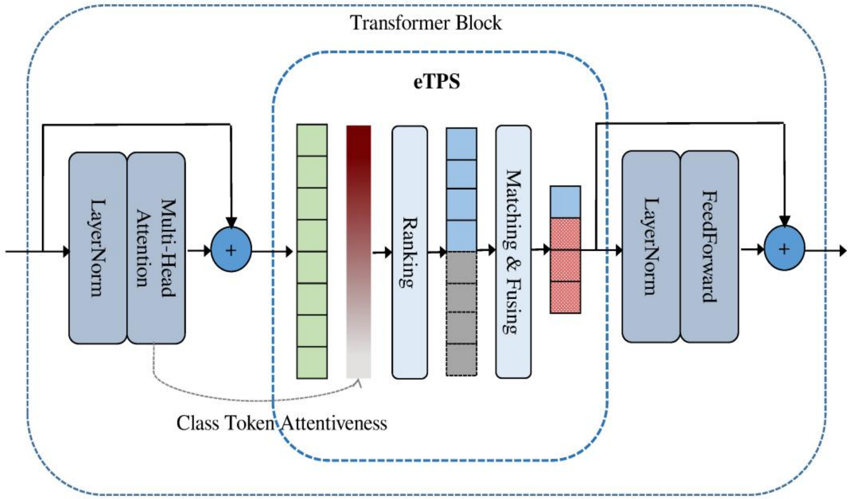
Quick Preview

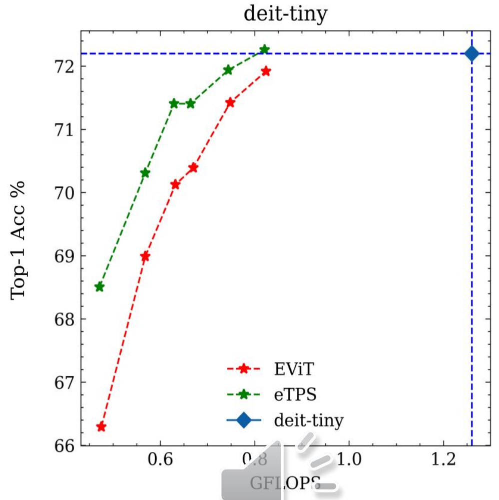
- Two flexible variants: the inter-block version dTPS and the intra-block version eTPS, which are plug-and-play blocks for both **vanilla ViTs** and **hybrid ViTs.**

- dTPS and eTPS surpass baselines dynamicViT and EViT by a large margin**.**





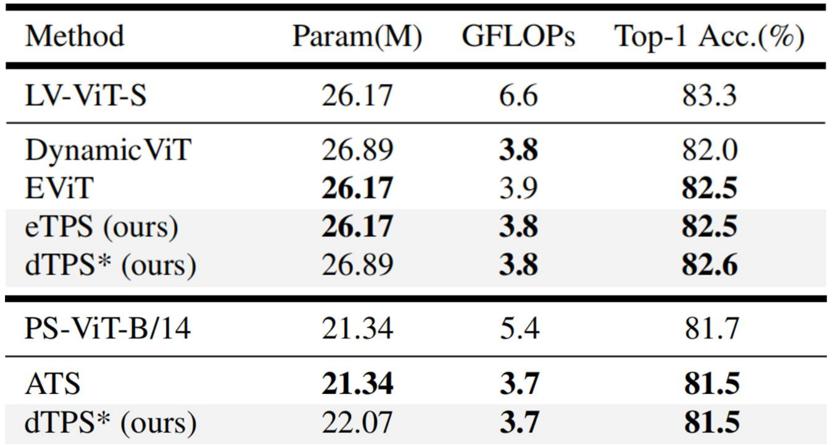




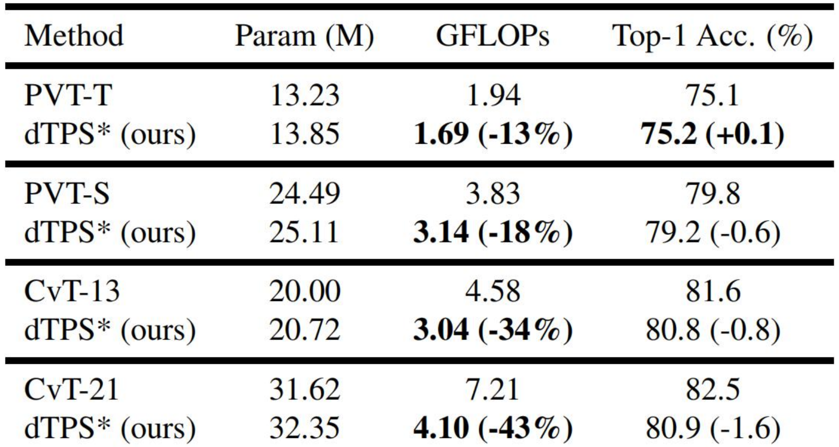
Quick Preview

- TPS can be extended to more vanilla ViTs…

…and hybrid ViTs.

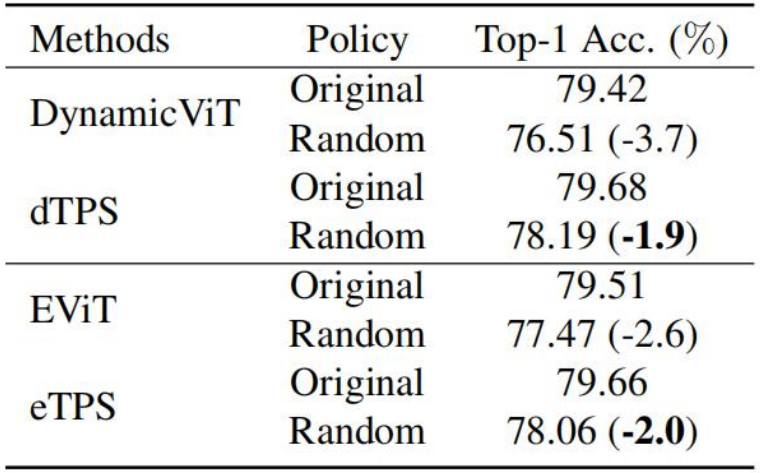


Method	Param(M)	GFLOPs	Top-1 Acc.(%)
LV-ViT-S	26.17	6.6	83.3
DynamicViT	26.89	**3.8**	82.0
EViT	**26.17**	3.9	**82.5**
eTPS (ours)	**26.17**	**3.8**	**82.5**
dTPS* (ours)	26.89	**3.8**	**82.6**
PS-ViT-B/14	21.34	5.4	81.7
ATS	**21.34**	**3.7**	**81.5**
dTPS* (ours)	22.07	**3.7**	**81.5**



Method	Param (M)	GFLOPs	Top-1 Acc. (%)
PVT-T	13.23	1.94	75.1
dTPS* (ours)	13.85	**1.69 (-13%)**	**75.2 (+0.1)**
PVT-S	24.49	3.83	79.8
dTPS* (ours)	25.11	**3.14 (-18%)**	79.2 (-0.6)
CvT-13	20.00	4.58	81.6
dTPS* (ours)	20.72	**3.04 (-34%)**	80.8 (-0.8)
CvT-21	31.62	7.21	82.5
dTPS* (ours)	32.35	**4.10 (-43%)**	80.9 (-1.6)

- Compared with previous methods, our TPS demonstrates robustness under random policies.



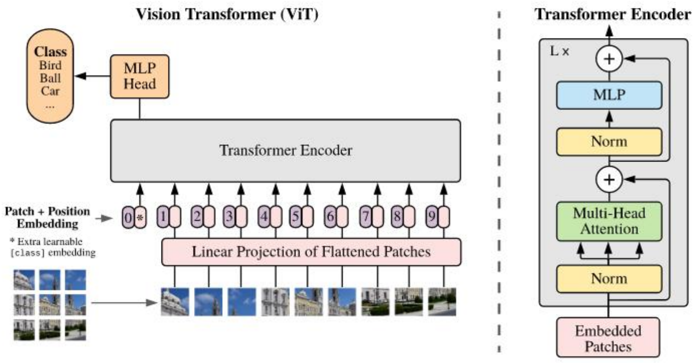
Methods	Policy	Top-1 Acc. (%)
DynamicViT	Original	79.42
	Random	76.51 (-3.7)
dTPS	Original	79.68
	Random	78.19 (**-1.9**)
EViT	Original	79.51
	Random	77.47 (-2.6)
eTPS	Original	79.66
	Random	78.06 (**-2.0**)

Motivation

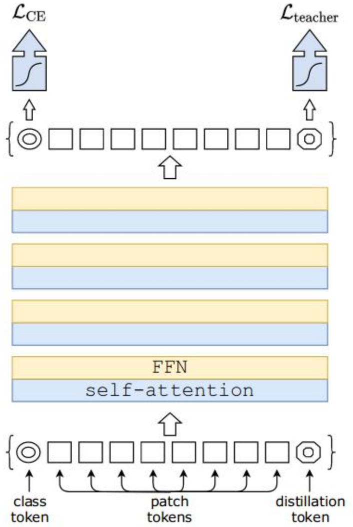
- Vision Transformer: new arch from NLP
- Strong performance but high computation cost

Vanilla ViTs

ViT (ICLR 2021)

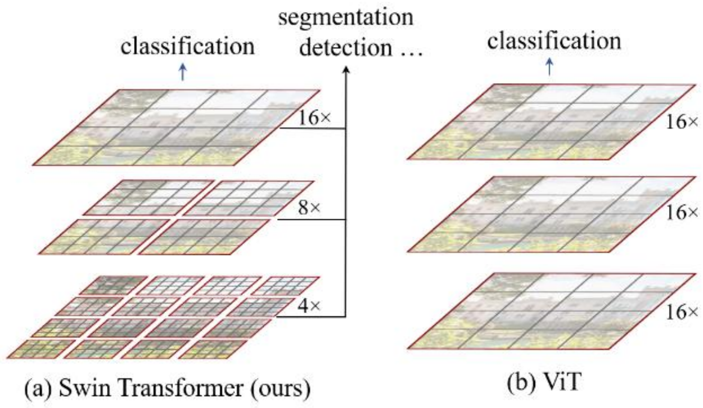


DeiT (ICML2021)



Hybrid ViTs

Swin (ICCV2021)



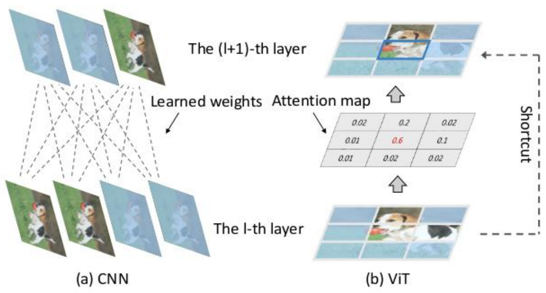
(a) Swin Transformer (ours)

(b) ViT

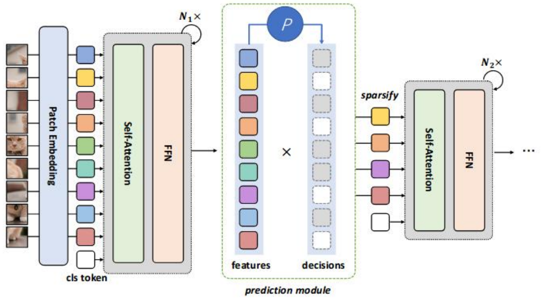
Motivation

Speed up ViTs from the perspective of token redundancy

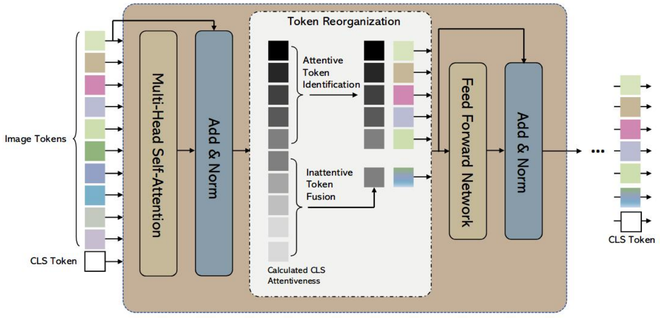
PatchSlimming (CVPR2022)



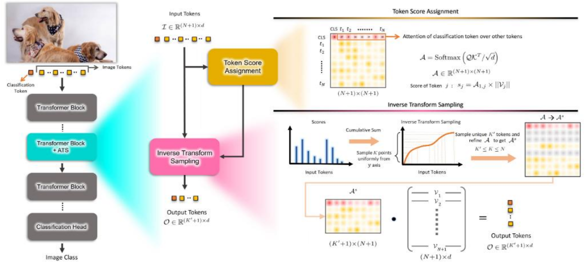
DynamicViT (NeurIPS2021)



EViT (ICLR2022)



ATS (ECCV2022)

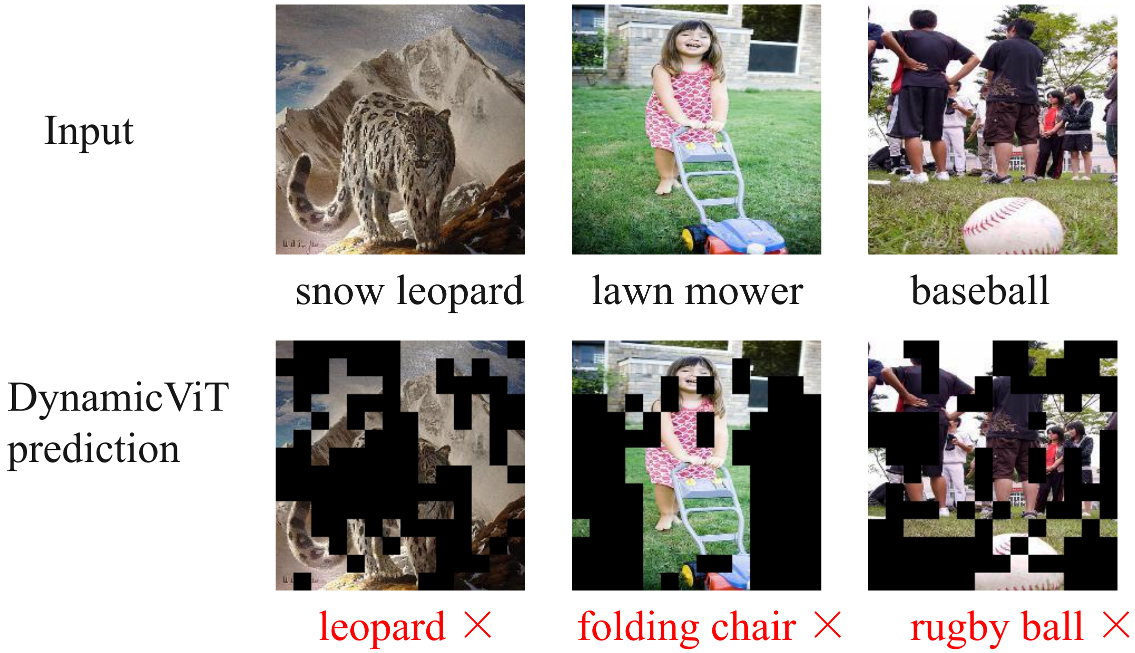


Limitations of Prior Works

- context information loss
- extra package tokens in EViT, SPViT
- non-constant-shape models
- complex training techniques

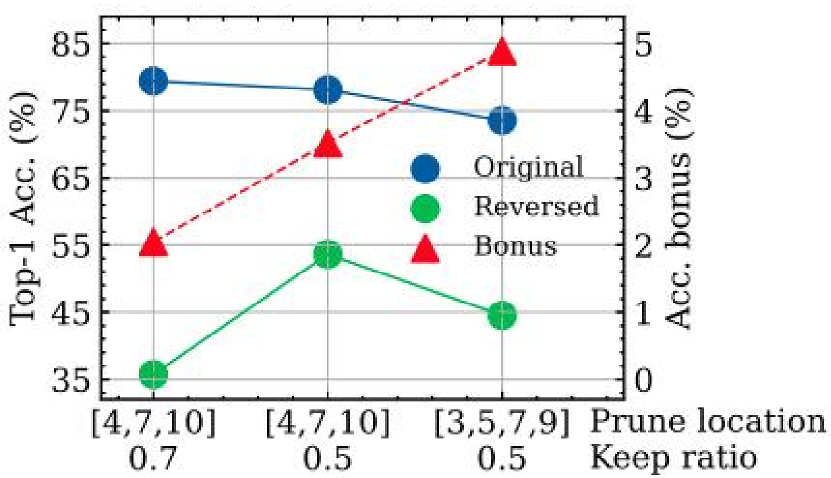
Introduction

Wrong predictions led by pruning



Toy Experiments

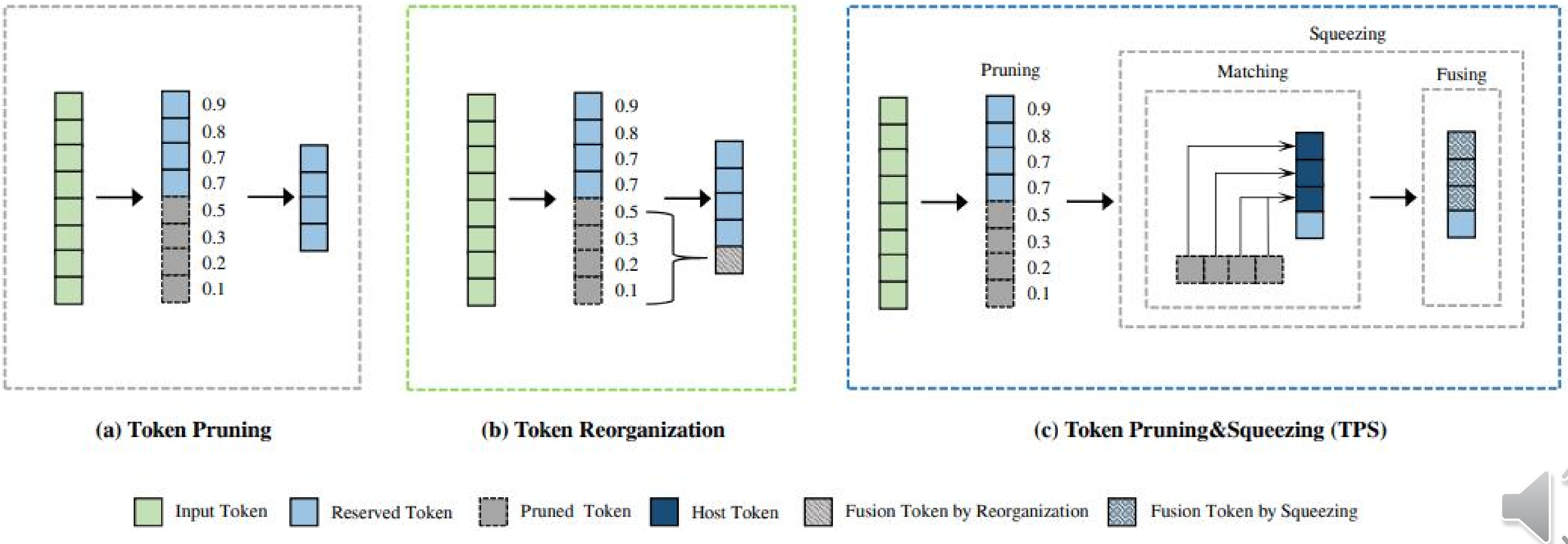
Bonus accuracy from pruned tokens increases along with more aggressive pruning strategies.



Introduction

TPS: Joint Token Pruning and Squeezing

1. preserve information from pruned tokens
2. constant-shape
3. no extra tokens



(a) Token Pruning

(b) Token Reorganization

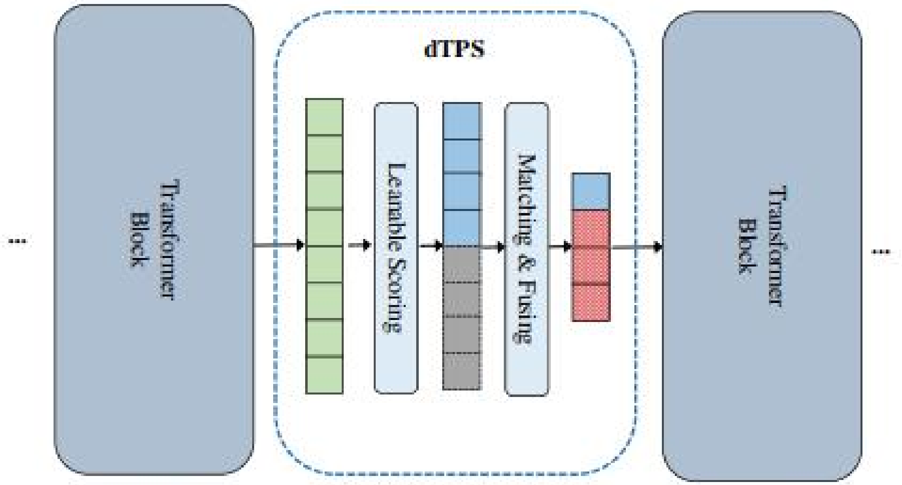
(c) Token Pruning&Squeezing (TPS)

Method

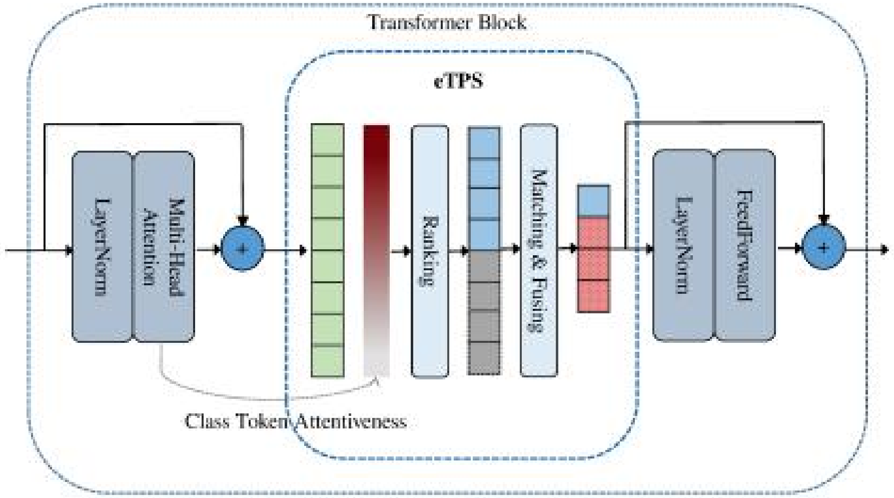
Step1: **Pruning**：

Two variants for covering both inter-block & intra-block pruning

- **dTPS** vs DynamicViT: learnable scoring, inter-block
- **eTPS** vs EViT: attention scoring, intra-block



(a) The inter-block variant of our TPS: dTPS.

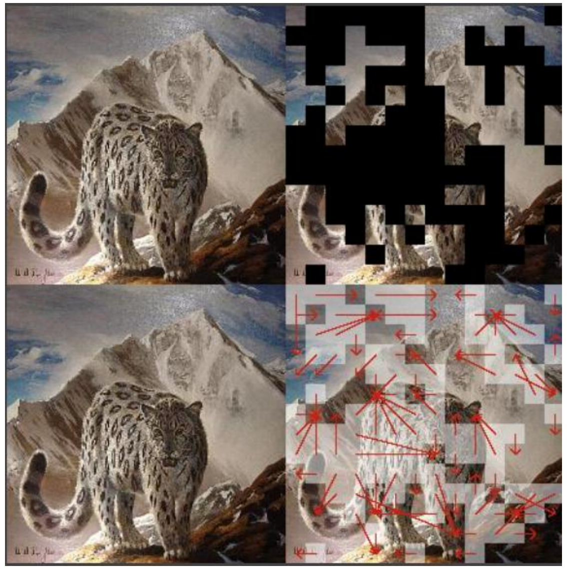


(b) The intra-block variant of our TPS: eTPS.

Method

Step2: **Squeezing**：Matching + Fusing

- **Matching**:
  - a unidirectional nearest-neighbor matching algorithm from pruned set to reserved set in a many-to-one manner
  - derive the matching relations based on a similarity matrix (cosine > previous attention)
- **Fusing**:
  - Similarity-based weighting, implementation with regular operations



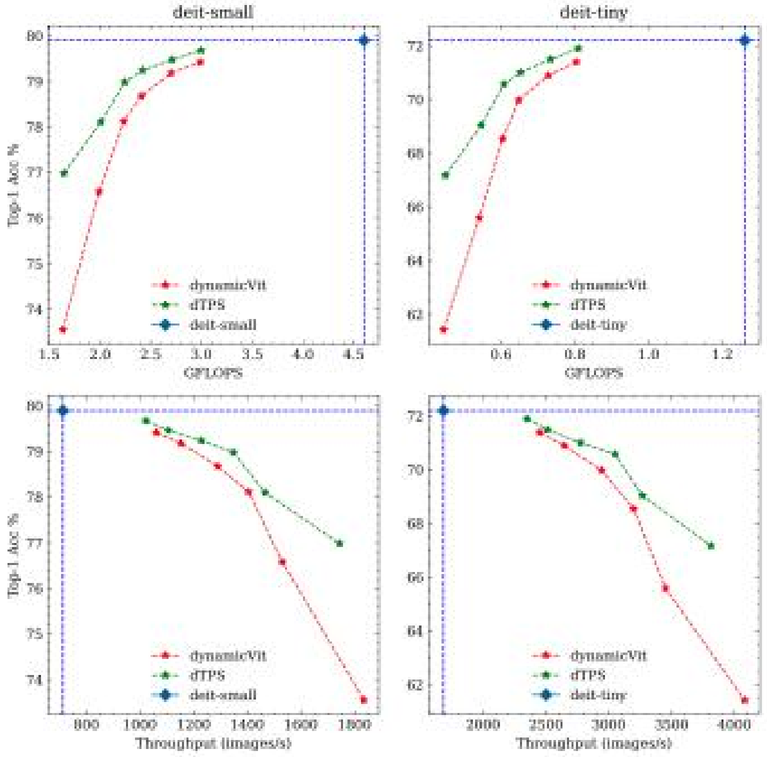
$$\boldsymbol{y}_j = w_j \boldsymbol{x}_j + \sum_{\boldsymbol{x}_i \in S^p} w_i \boldsymbol{x}_i, \tag{4}$$

$$w_i = \frac{\exp(c_{i,j}) m_{i,j}}{\sum_{\boldsymbol{x}_i \in S^p} \exp(c_{i,j}) m_{i,j} + \mathrm{e}}. \tag{5}$$

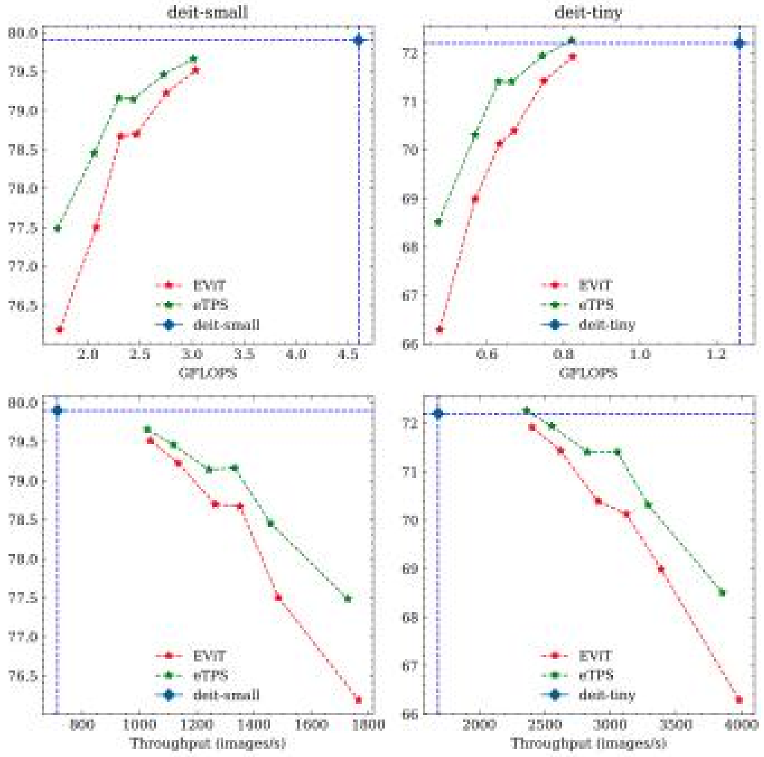
$$w_j = \frac{\mathrm{e}}{\sum_{\boldsymbol{x}_i \in S^p} \exp(c_{i,j}) m_{i,j} + \mathrm{e}}. \tag{6}$$

Experiments

Main Results



(a) Comparison between our dTPS and dynamicViT on DeiT.

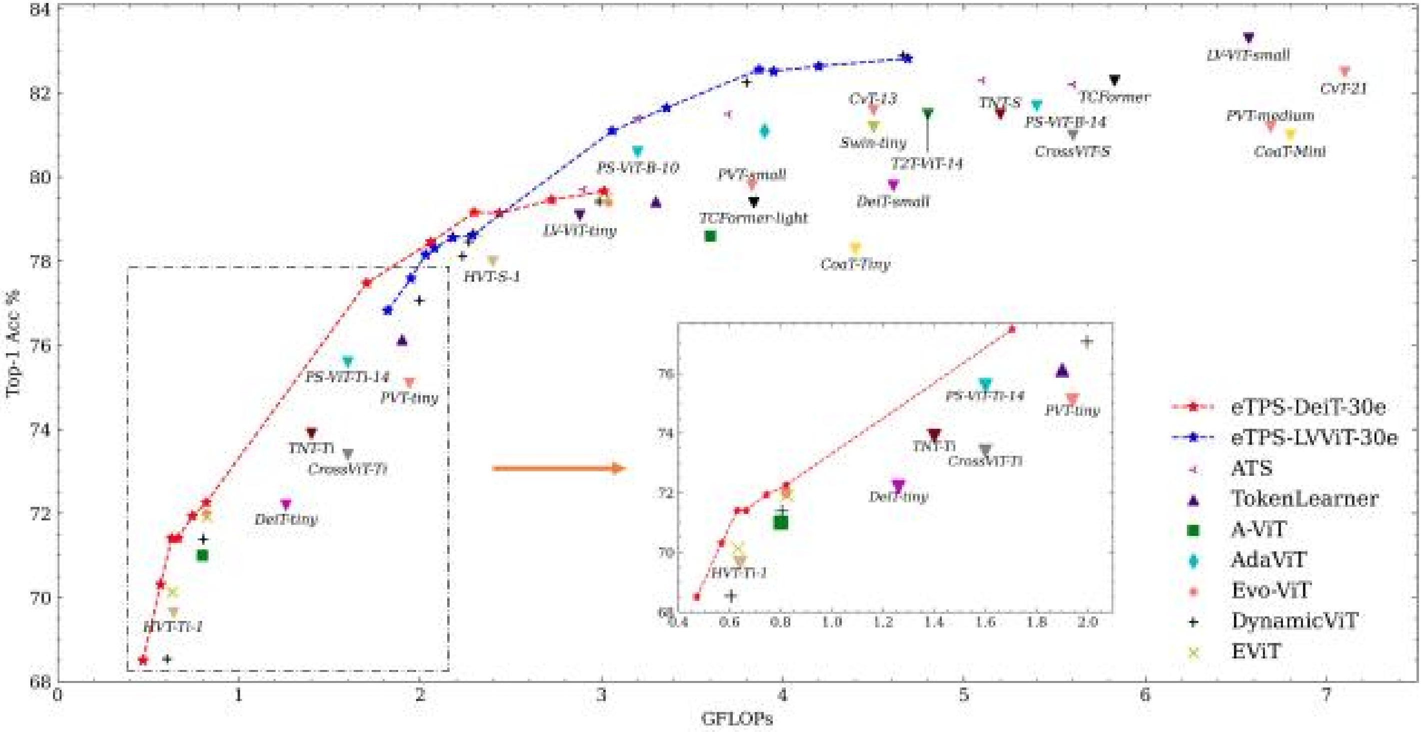


(b) Comparison between our eTPS and EViT on DeiT.

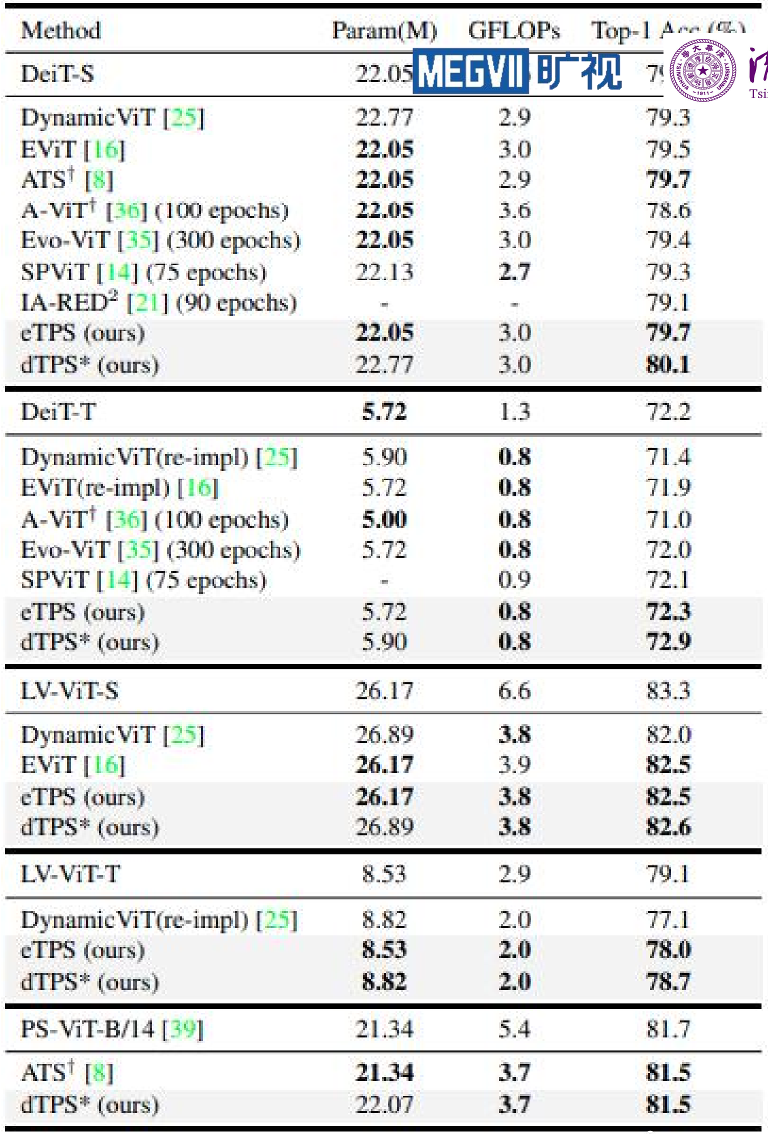
Comparison to baselines

Experiments

Main Results



Comparison to current SOTAs

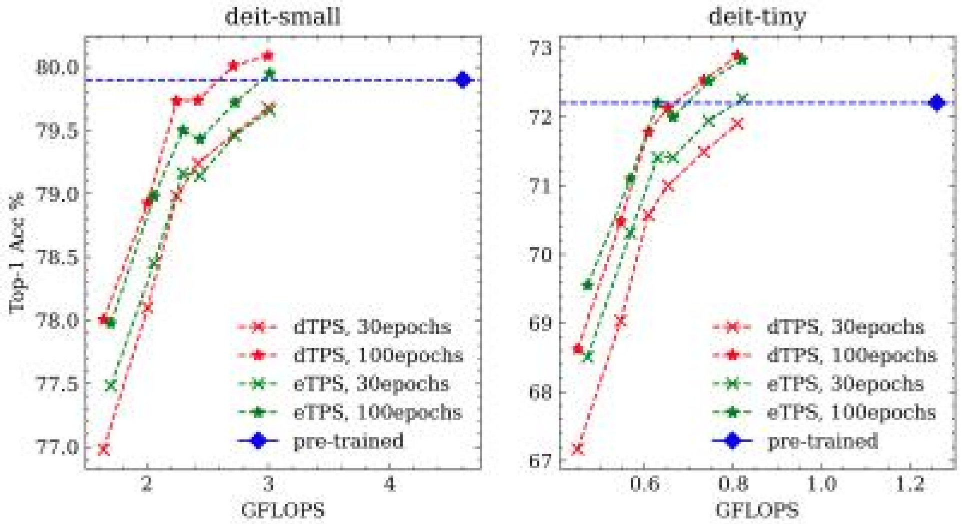


Method	Param(M)	GFLOPs	Top-1 Acc (%)
DeiT-S	22.05	[illegible]	[illegible]
DynamicViT [25]	22.77	2.9	79.3
EViT [16]	**22.05**	3.0	79.5
ATS† [8]	**22.05**	2.9	**79.7**
A-ViT† [36] (100 epochs)	**22.05**	3.6	78.6
Evo-ViT [35] (300 epochs)	**22.05**	3.0	79.4
SPViT [14] (75 epochs)	22.13	**2.7**	79.3
IA-RED^2 [21] (90 epochs)	-	-	79.1
eTPS (ours)	**22.05**	3.0	**79.7**
dTPS* (ours)	22.77	3.0	**80.1**
DeiT-T	**5.72**	1.3	72.2
DynamicViT(re-impl) [25]	5.90	**0.8**	71.4
EViT(re-impl) [16]	5.72	**0.8**	71.9
A-ViT† [36] (100 epochs)	**5.00**	**0.8**	71.0
Evo-ViT [35] (300 epochs)	5.72	**0.8**	72.0
SPViT [14] (75 epochs)	-	0.9	72.1
eTPS (ours)	5.72	**0.8**	**72.3**
dTPS* (ours)	5.90	**0.8**	**72.9**
LV-ViT-S	26.17	6.6	83.3
DynamicViT [25]	26.89	**3.8**	82.0
EViT [16]	**26.17**	3.9	**82.5**
eTPS (ours)	**26.17**	**3.8**	**82.5**
dTPS* (ours)	26.89	**3.8**	**82.6**
LV-ViT-T	8.53	2.9	79.1
DynamicViT(re-impl) [25]	8.82	2.0	77.1
eTPS (ours)	**8.53**	**2.0**	**78.0**
dTPS* (ours)	**8.82**	**2.0**	**78.7**
PS-ViT-B/14 [39]	21.34	5.4	81.7
ATS† [8]	**21.34**	**3.7**	**81.5**
dTPS* (ours)	22.07	**3.7**	**81.5**

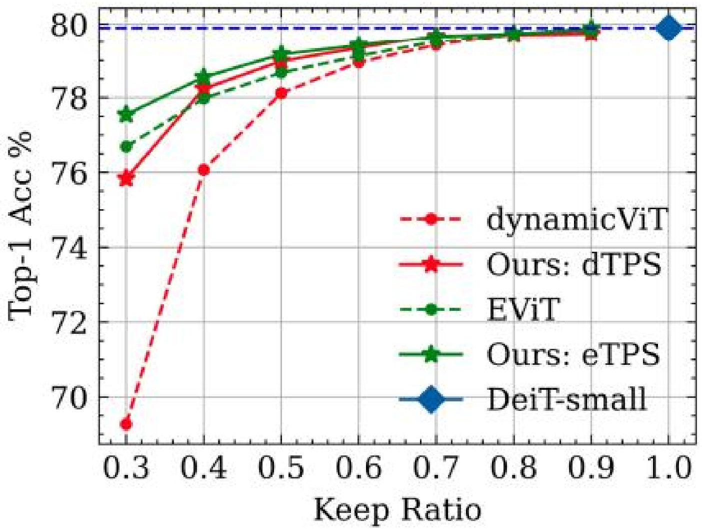
Extension on more backbones

Experiments

Ablation Study



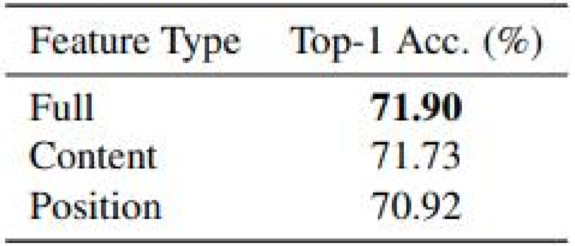
Epochs of training



Different keeping ratios

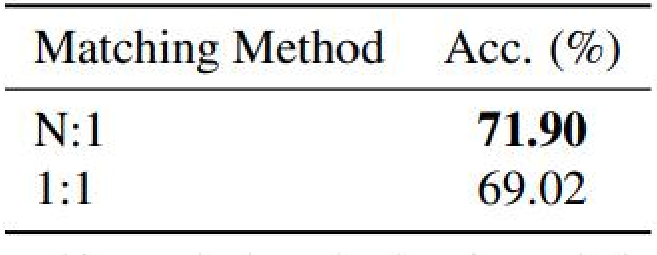
Experiments

Ablation Study



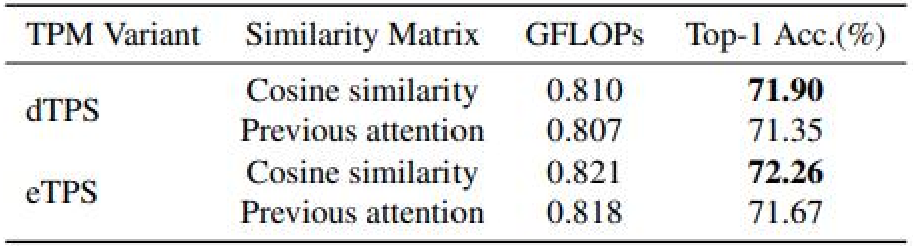
Feature Type	Top-1 Acc. (%)
Full	**71.90**
Content	71.73
Position	70.92

Feature type used in matching



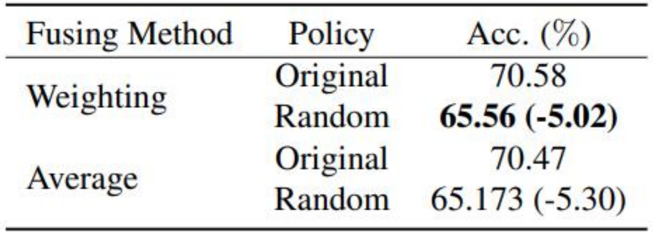
Matching Method	Acc. (%)
N:1	**71.90**
1:1	69.02

Differen matching methods



TPM Variant	Similarity Matrix	GFLOPs	Top-1 Acc.(%)
dTPS	Cosine similarity	0.810	**71.90**
	Previous attention	0.807	71.35
eTPS	Cosine similarity	0.821	**72.26**
	Previous attention	0.818	71.67

Different similarity matrix

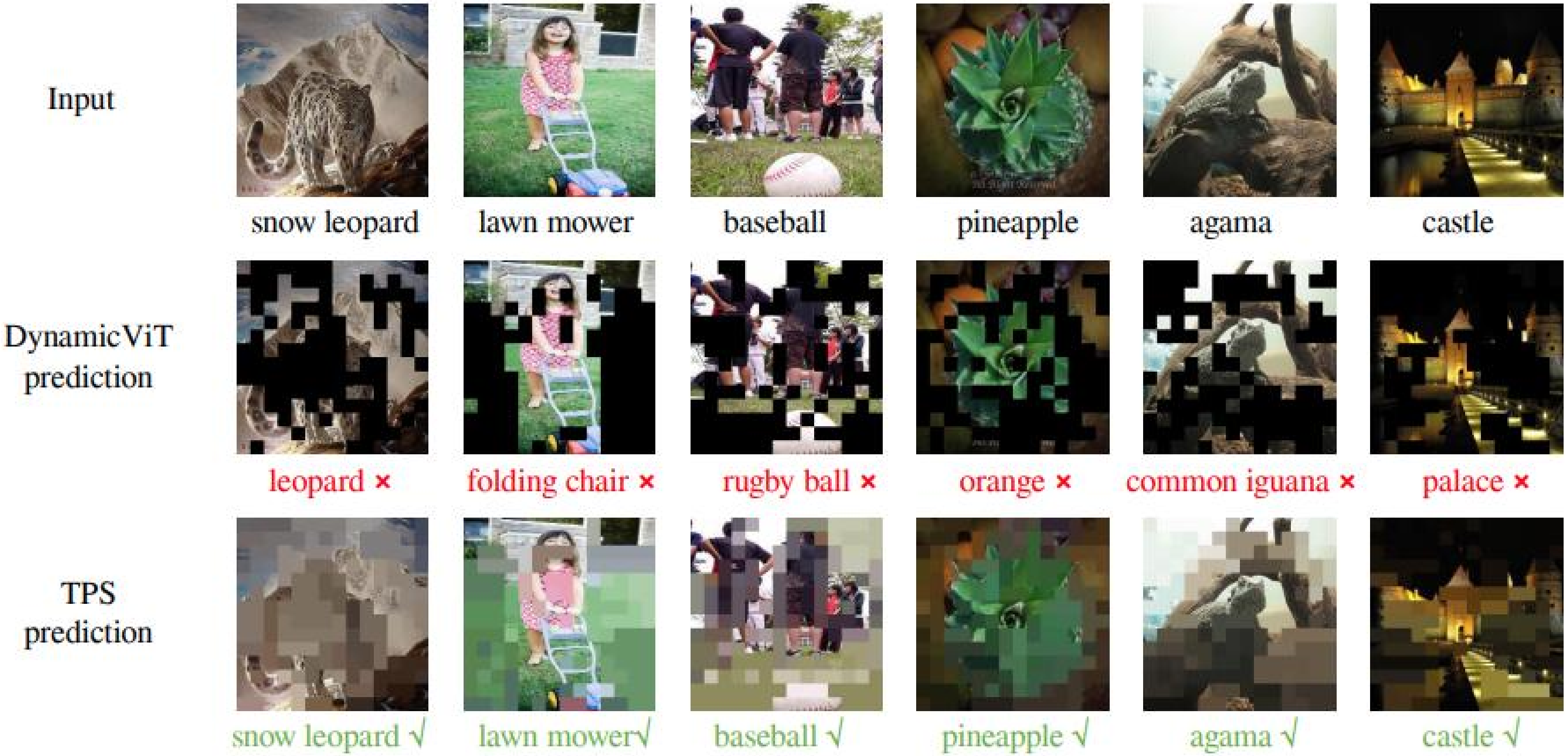


Fusing Method	Policy	Acc. (%)
Weighting	Original	70.58
	Random	**65.56 (-5.02)**
Average	Original	70.47
	Random	65.173 (-5.30)

Different fusing methods

Experiments

More Visualizations



Thanks