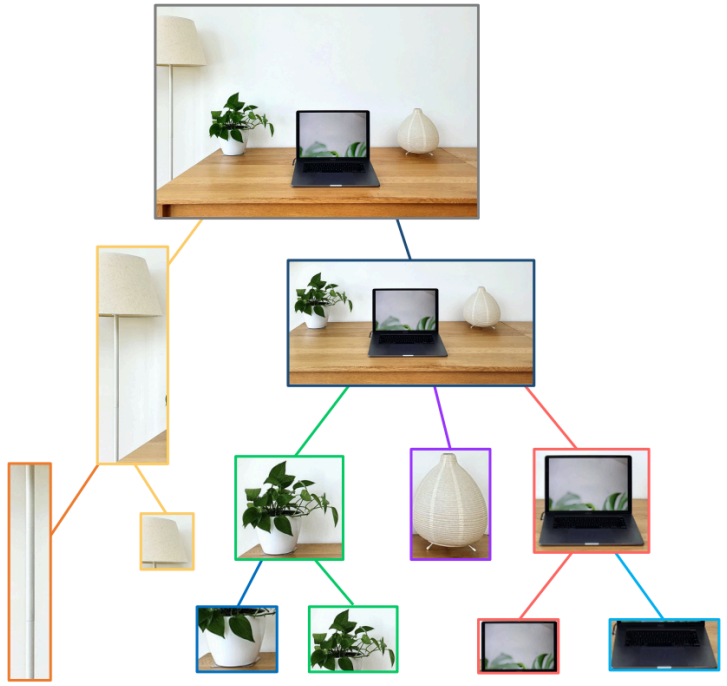
Visual Dependency Transformers: Dependency Tree Emerges from Reversed Attention

Mingyu Ding · Yikang Shen · Lijie Fan · Zhenfang Chen · Zitian Chen · Ping Luo · Joshua B. Tenenbaum · Chuang Gan

UC Berkeley HKU MIT MIT-IBM Watson AI Lab UMass Amherst

Background

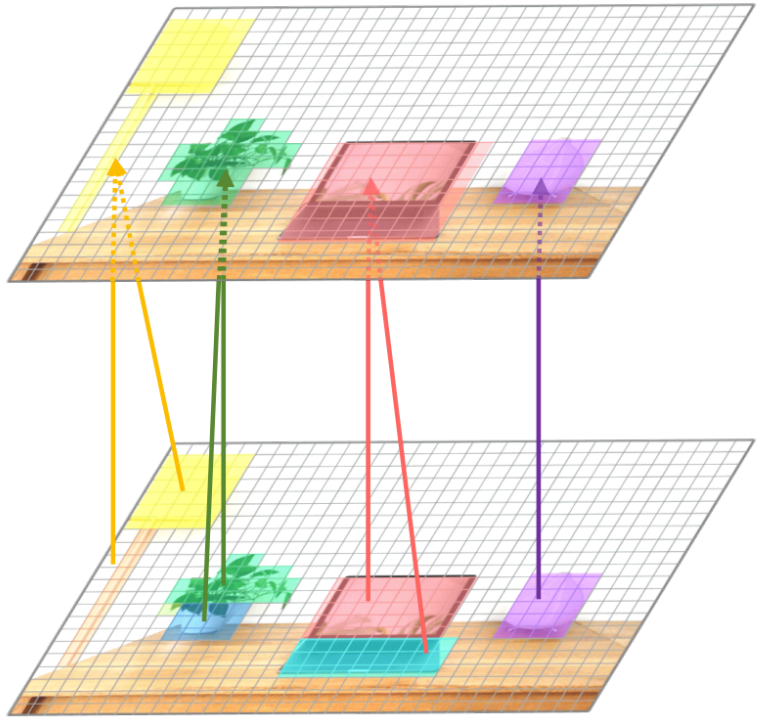
- Humans possess a versatile mechanism for extracting structured representations of our visual world.
- When looking at an image, we can decompose the scene into entities and their parts as well as obtain the dependencies between them.
- To mimic such capability, we propose Visual Dependency Transformers (DependencyViT) that can induce visual dependencies without any labels.



Method: Reversed Attention

- A child token in reversed attention is trained to attend to its parent tokens and send information following a normalized probability distribution rather than gathering information in conventional self-attention.

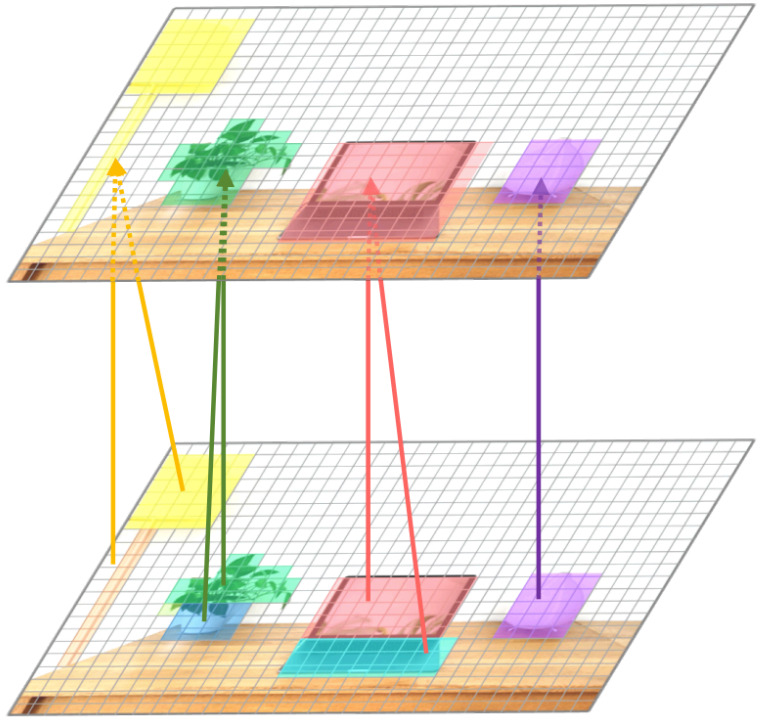
- Hierarchies naturally emerge from reversed attention layers, and a dependency tree is progressively induced from leaf nodes to the root node unsupervisedly.



Method: Reversed Attention

- Entities and their parts in an image are represented by different subtrees, enabling part partitioning from dependencies.

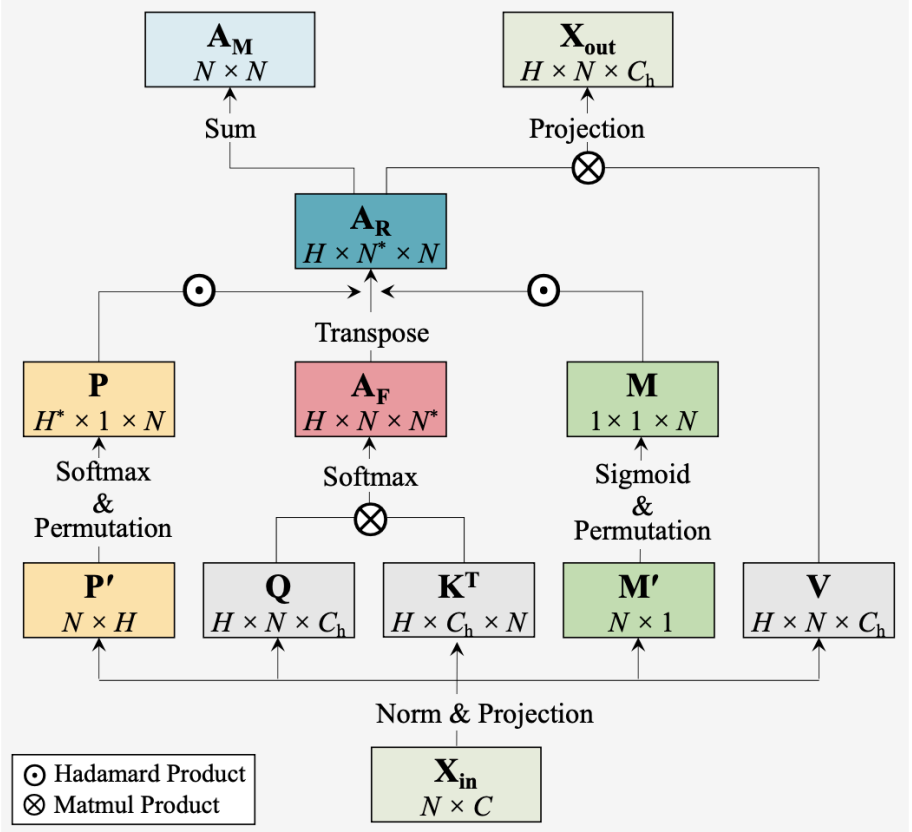
- Dynamic visual pooling is made possible. The leaf nodes which rarely send message can be pruned without hindering performance.



Dependency Block

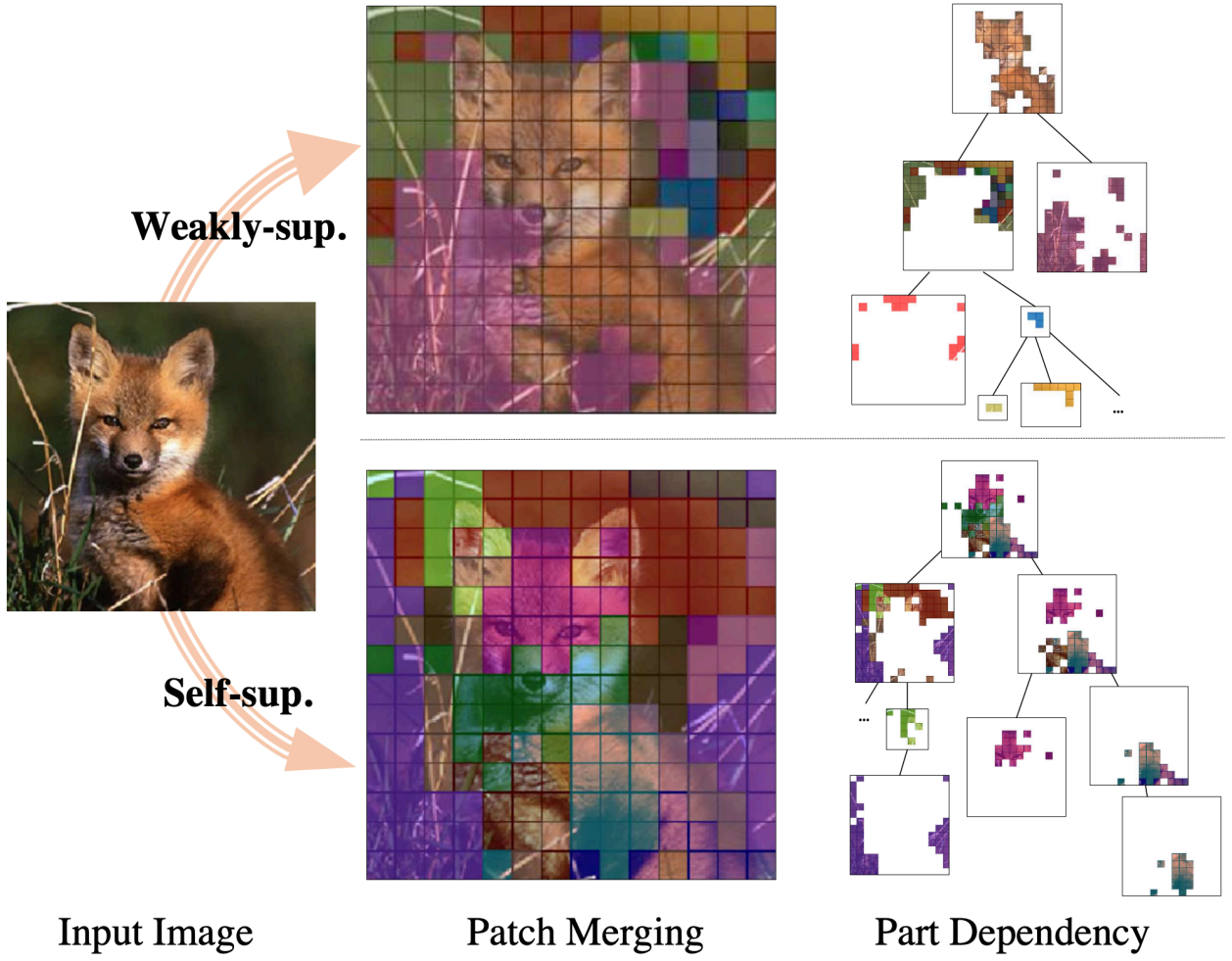
- We obtain the reversed attention matrices A_R by transposing the forward attention weights A_F with a head selector P and a message controller M imposing on it.

- H: number of heads
- N: number of tokens
- C: token dim

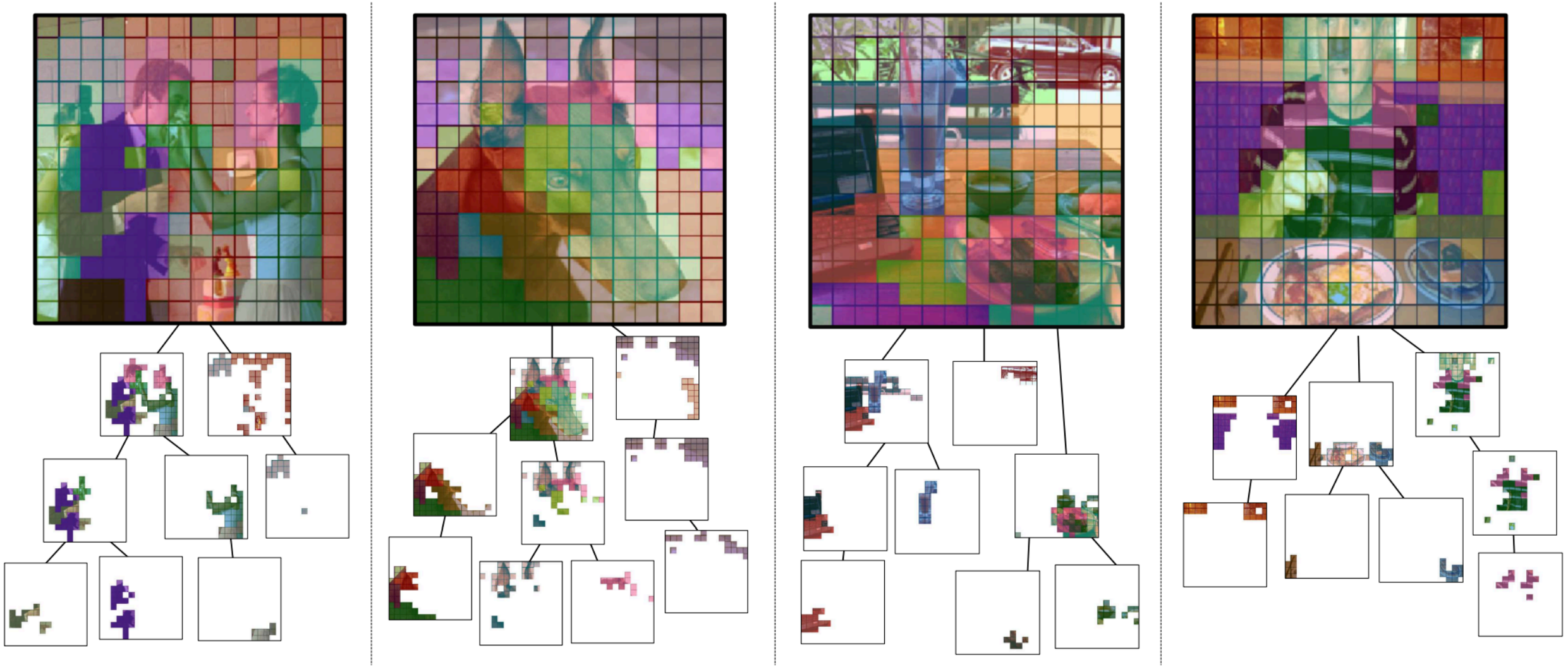


Learning paradigm

- Weakly-supervised pretrained model focuses more on the entire object.
- self-supervised pretrained model can capture more fine-grained partaware dependencies.
- The parsed dependency tree is expected to help many downstream tasks, such as saliency detection and part segmentation.



Visualizations



Experiments

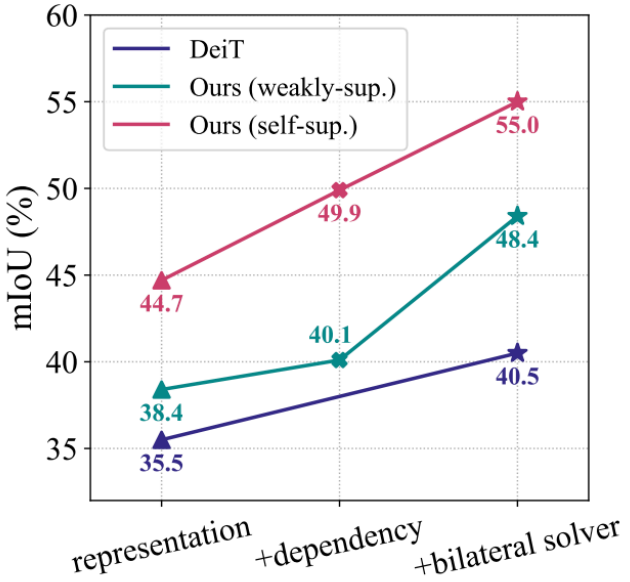
- Part Segmentation

Method	Pretraining Type	Part Discovery by	Pascal-Part [8]		Car-Parts [52]	
			mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)
DeiT [64]	weakly-sup.	clustering	7.2	22.6	8.9	29.5
DeiT [64]	weakly-sup.	maximum spanning	18.9	35.5	17.8	37.7
DependencyViT	weakly-sup.	clustering	11.6	31.7	10.9	29.7
DependencyViT	weakly-sup.	maximum spanning	23.2	41.7	22.6	40.0
DependencyViT-Lite	weakly-sup.	dynamic pruning	25.7	44.0	24.3	44.9
iBOT [93]	self-sup.	maximum spanning	25.1	44.8	25.7	46.1
DependencyViT	self-sup.	maximum spanning	**31.2**	**51.1**	**30.6**	**50.3**

Experiments

- Unsupervised Saliency Detection

Method	ECSSD [59]			DUTS [69]			DUT-OMRON [83]		
	$maxF_\beta$ (%)	IoU (%)	Acc. (%)	$maxF_\beta$ (%)	IoU (%)	Acc. (%)	$maxF_\beta$ (%)	IoU (%)	Acc. (%)
DeepUSPS [50]	58.4	44.0	79.5	42.5	30.5	77.3	41.4	30.5	77.9
HS [82]	67.3	50.8	84.7	50.4	36.9	82.6	56.1	43.3	84.3
wCtr [95]	68.4	51.7	86.2	52.2	39.2	83.5	54.1	41.6	83.8
WSC [41]	68.3	49.8	85.2	52.8	38.4	86.2	52.3	38.7	86.5
DeiT-Tiny [64]	49.3	40.5	72.7	34.2	26.8	72.7	33.2	27.2	71.1
DependencyViT-T (self-sup.)	62.1	55.0	78.4	43.0	35.9	73.2	32.5	28.0	67.2
DependencyViT-T (weakly-sup.)	62.0	48.4	83.6	53.8	37.0	87.5	52.0	39.7	**88.4**
Deit-Small [64]	56.3	49.9	76.8	38.9	33.3	72.2	34.5	30.7	71.2
DependencyViT-S (self-sup.)	65.2	**58.4**	80.2	46.4	39.1	74.8	34.7	30.4	68.3
DependencyViT-S (weakly-sup.)	**70.5**	53.4	**86.8**	**57.9**	**44.5**	**87.8**	**53.9**	**44.2**	**88.4**



Experiments

- Image Classification

Model	Direction	Head Selector	Message Controller	#Params (M)	FLOPs (G)	Top-1 (%)
Baseline (DeiT) [64]	forward	×	×	5.7	1.3	73.3
Forward + P	forward	√	×	5.7	1.3	73.4
Forward + M	forward	×	√	6.1	1.3	74.8
Forward + P + M	forward	√	√	6.2	1.3	74.8
Reverse + P	reverse	√	×	5.7	1.3	73.6
Reverse + M	reverse	×	√	6.1	1.3	74.9
Reverse + P + M (DependencyViT)	reverse	√	√	6.2	1.3	**75.4**
DependencyViT-Lite (forward)	forward	√	√	6.2	0.8	71.1
DependencyViT-Lite (reverse)	reverse	√	√	6.2	0.8	73.7

Model	Hierarchical	Cost	#Params (M)	FLOPs (G)	Top-1 (%)
ResNet-18 [27]	√	low	11.7	1.8	69.9
ConvMixer-512/16 [65]	×	high	5.4	–	73.7
DeiT-Tiny/16 [64]	×	high	5.7	1.3	72.2
CrossViT-Tiny [6]	×	high	6.9	1.6	73.4
PVT-Tiny [71]	√	low	13.2	1.9	75.1
DependencyViT-Lite-T	×	low	6.2	0.8	73.7
DependencyViT-T	×	high	6.2	1.3	**75.4**
ResNet-50 [27]	√	low	25.0	4.1	76.2
ConvMixer-768/32 [65]	×	high	21.1	–	80.2
DeiT-Small/16 [64]	×	high	22.1	4.5	79.8
CrossViT-Small [6]	×	high	26.7	5.6	81.0
PVT-Small [71]	√	low	24.5	3.8	79.8
Swin-Tiny [48]	√	low	28.3	4.5	81.2
CvT-13 [75]	√	high	20.0	4.5	81.6
DynamicViT-LV-S/0.5 [53]*	×	–	26.9	3.7	82.0
PVTv2-B2 [70]	√	low	25.4	4.0	82.0
DependencyViT-Lite-S	×	low	24.0	3.0	80.6
DependencyViT-S	×	high	24.0	5.0	**82.1**

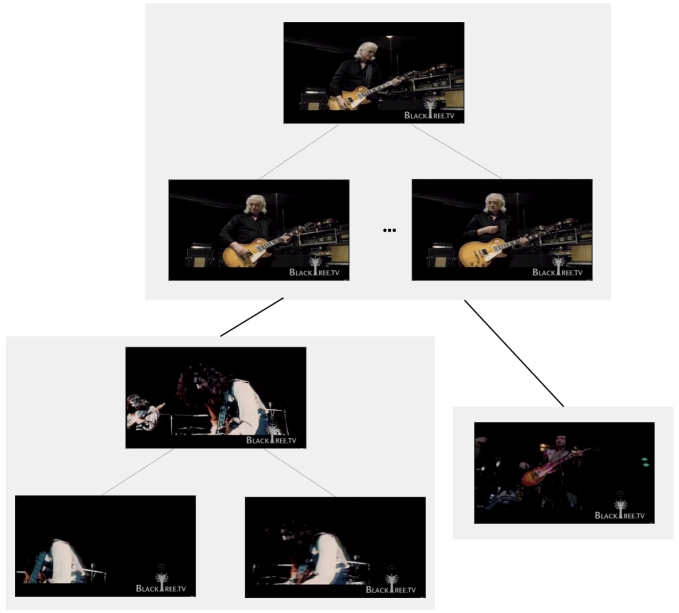
Experiments

- Dynamic Pruning

Model	kept tokens	#Params (M)	FLOPs (G)	Top-1 (%)
DependencyViT-Lite-32	32	6.2	0.6	72.4
DependencyViT-Lite-64	64	6.2	0.8	73.7
DependencyViT-Lite-128	128	6.2	1.0	74.9
DependencyViT	196	6.2	1.3	75.4

- Dependency in Videos

Method	Top-1 (%)	Top-5 (%)	FLOPs (G)	Frames	Resolution
TimeSformer	76.9	92.7	0.20	8	224
TimeSformer-Lite	70.6	89.3	0.08	8	224
TimeSformer-HR	78.1	93.3	1.70	16	448
TimeSformer-HR-Lite	73.1	90.4	0.67	16	448
TimeSformer-L	79.8	94.1	2.38	96	224
TimeSformer-L-Lite	74.1	91.3	0.61	96	224



Thanks