CompSys 2026

Orthogonal Compression for Edge Vision Transformers: Combining Recursive Weight-Sharing with Token Merging

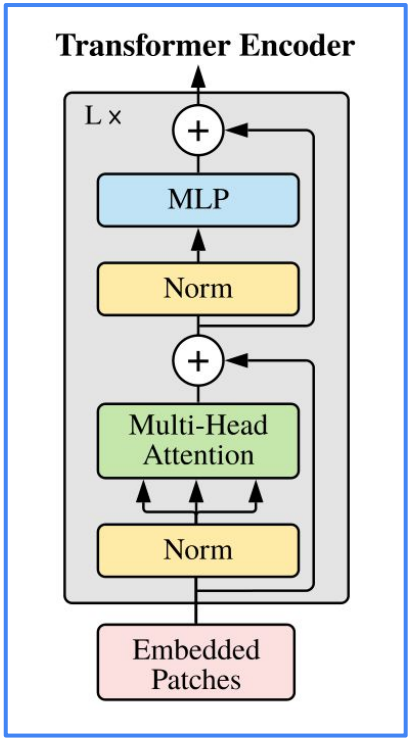
UNIVERSITY OF TWENTE.

Junseo Kim
Uraz Odyurt
Amirreza Yousefzadeh

Motivation

Vision Transformers [3]

- Large parameter counts
- Quadratic computational complexity



Standard ViT [3]

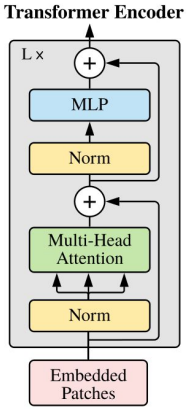
Motivation

Vision Transformers [3]

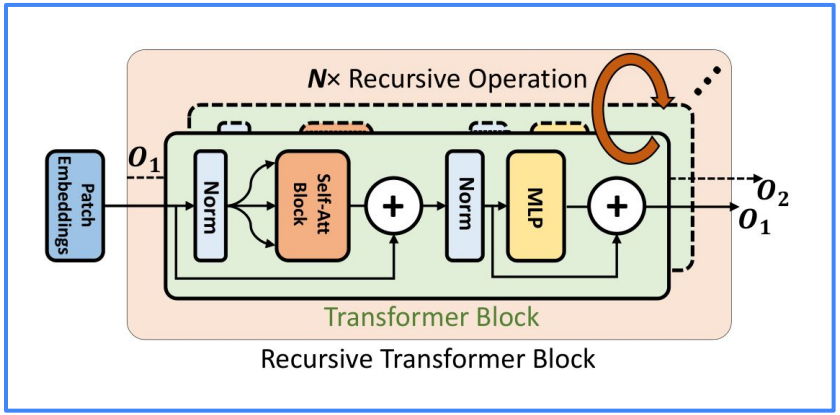
- Large parameter counts
- Quadratic computational complexity

Recursive Weight-Sharing [4]

- Reduces static parameter counts
- Adds computational complexity and memory



Standard ViT [3]



Recursive ViT [4]

Motivation

Vision Transformers [3]

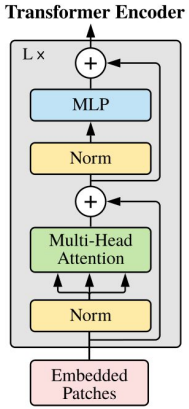
- Large parameter counts
- Quadratic computational complexity

Recursive Weight-Sharing [4]

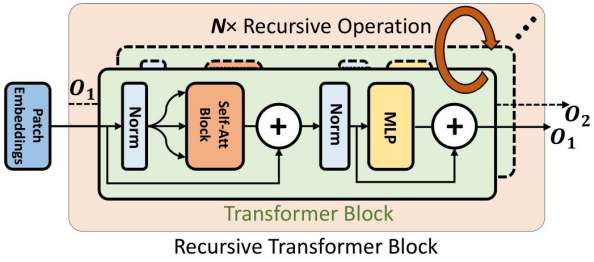
- Reduces static parameter counts
- Adds computational complexity and memory

Token Reduction Strategies [1]

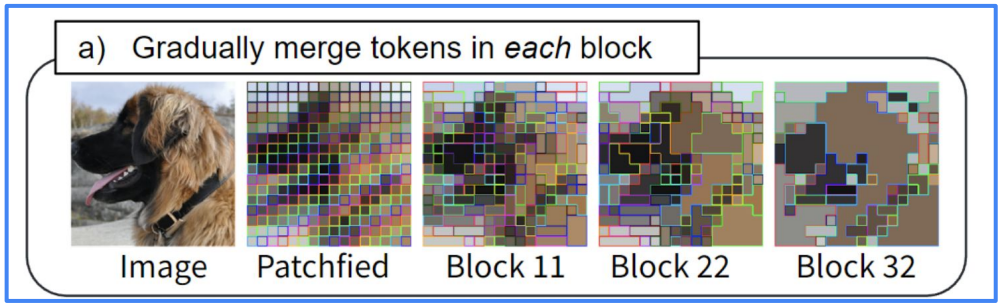
- **Reduce sequence lengths**
- **Reduce computational complexity and memory**



Standard ViT [3]



Recursive ViT [4]

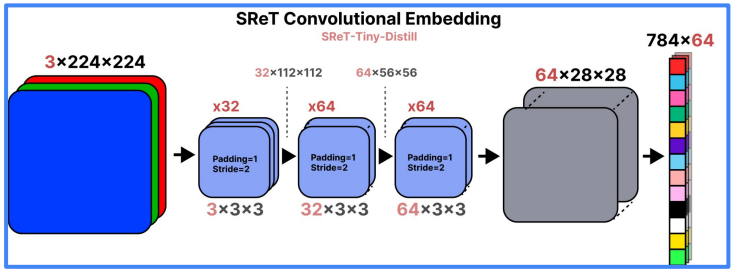


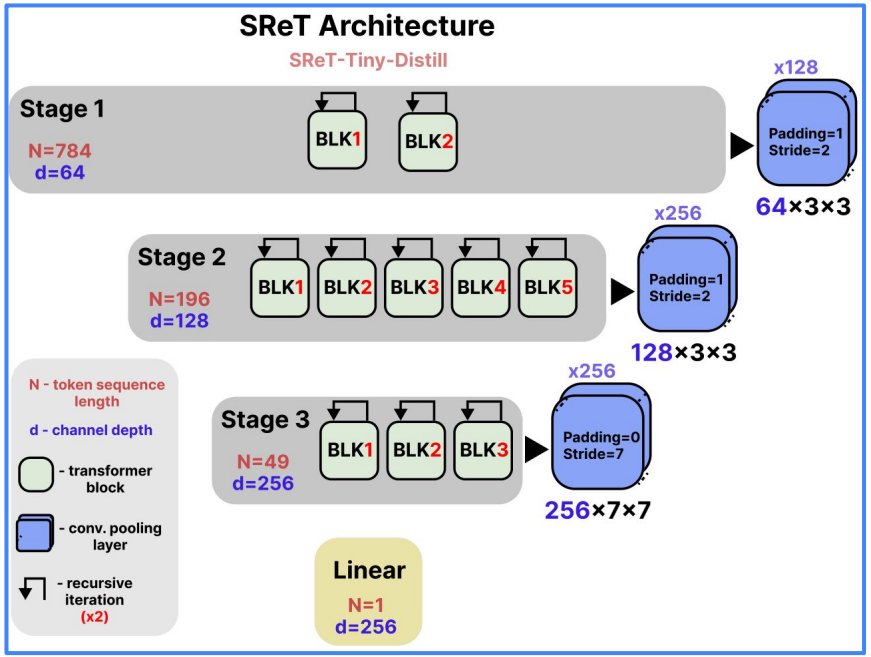
Token Merging [1]

Integration

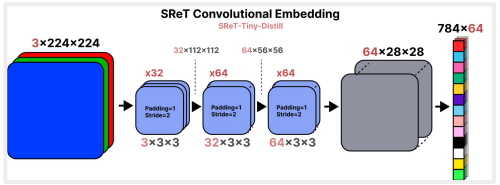
Sliced Recursive Transformer (SReT) [4]

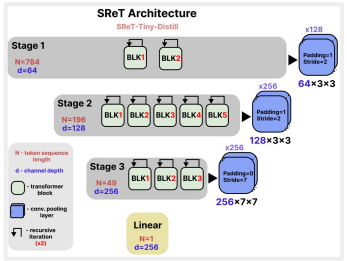
- Hierarchical architecture





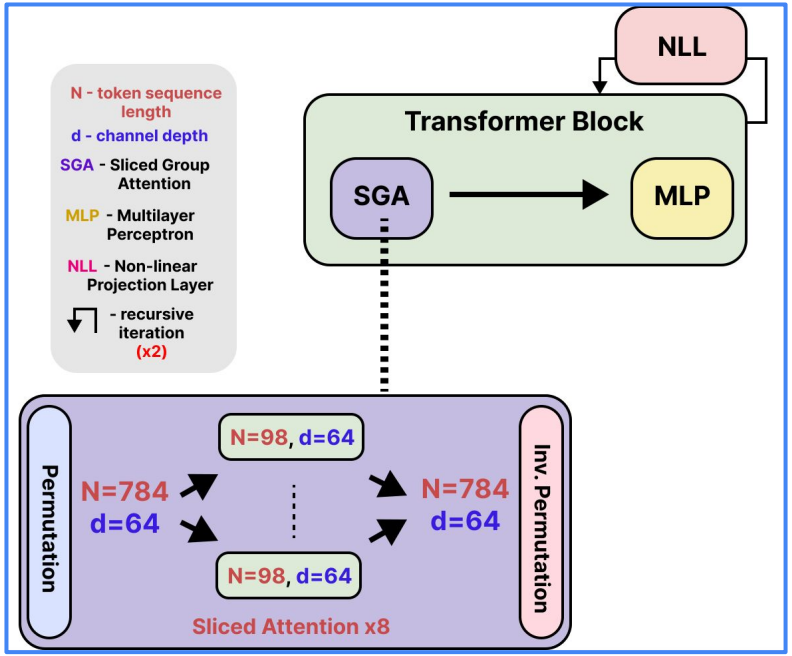
Integration





Sliced Recursive Transformer (SReT) [4]

- Hierarchical architecture
- Replaces Global Attention with Sliced Group Self-Attention (SGA)



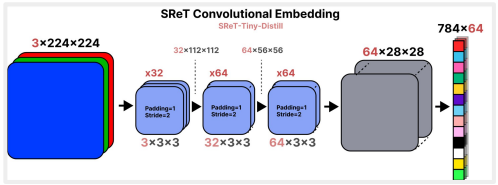
Integration

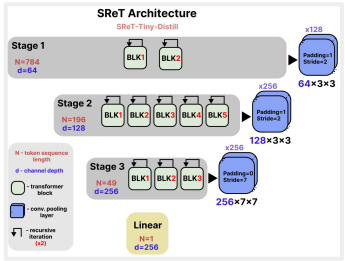
Sliced Recursive Transformer (SReT) [4]

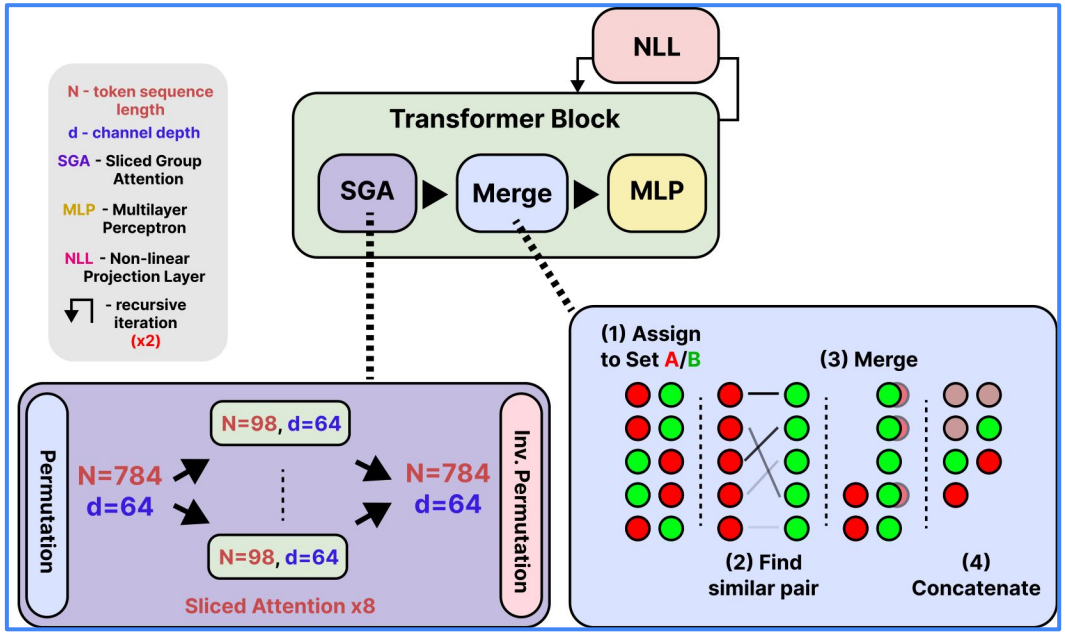
- Hierarchical architecture
- Replaces Global Attention with Sliced Group Self-Attention (SGA)

Token Merging (ToMe) [1]

- **Training-free token reduction strategy**
- **Bipartite Soft Matching (cosine similarity)**



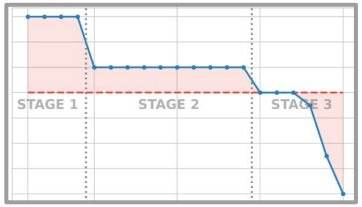




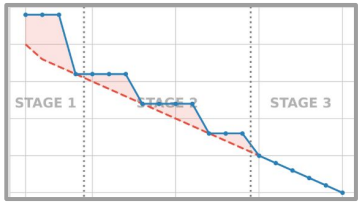
Experimental Setup

Best Reduction Schedule?

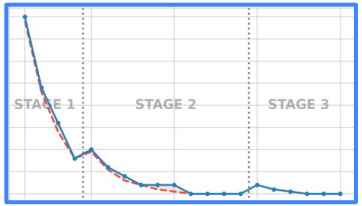
- Constant Reduction (ToMe) [1]



- Linear Reduction (ToMe) [1]



- **Exponential Reduction**



Experimental Setup

Best Reduction Schedule?

- Constant Reduction (ToMe) [1]
- Linear Reduction (ToMe) [1]
- Exponential Reduction

Evaluation Against Baseline?

- **Dataset: ImageNet-1K Validation Set [2]**
- **GPU: NVIDIA RTX 4060 Ti**
- **CPU: Intel Core Ultra 9 285K**

Baseline and Metrics

Model	Acc (%)	Params (M)	FLOPs (G)	Throughput (img/s) [GPU | BS=128]	Peak Memory (MB) [GPU | BS=128]
DeiT-Tiny-Distill [5]	74.40	5.91	2.17	1825.88	227.20
SReT-Tiny-Distill [4]	**77.42**	4.76	1.91	**1072.86**	**795.76**

Results - Constant Reduction

Merge a fixed number **(constant coefficient)** at every layer

Model	Acc (%)	Params (M)	FLOPs (G)	Throughput (img/s) [GPU | BS=128]	Peak Memory (MB) [GPU | BS=128]
SReT-Tiny-Distill [4]	77.42	4.76	1.91	1072.86	795.76
SReT-Tiny-Distill+ToMe **(r_const=10)**	71.01 -6.41	-	1.32 -30.9%	1176.27 +9.6%	769.79 -3.3%

Results - Linear Reduction

Merge **2 x (linear coefficient)** at first layer, then linearly decrease the merge rate down to **0** by last layer

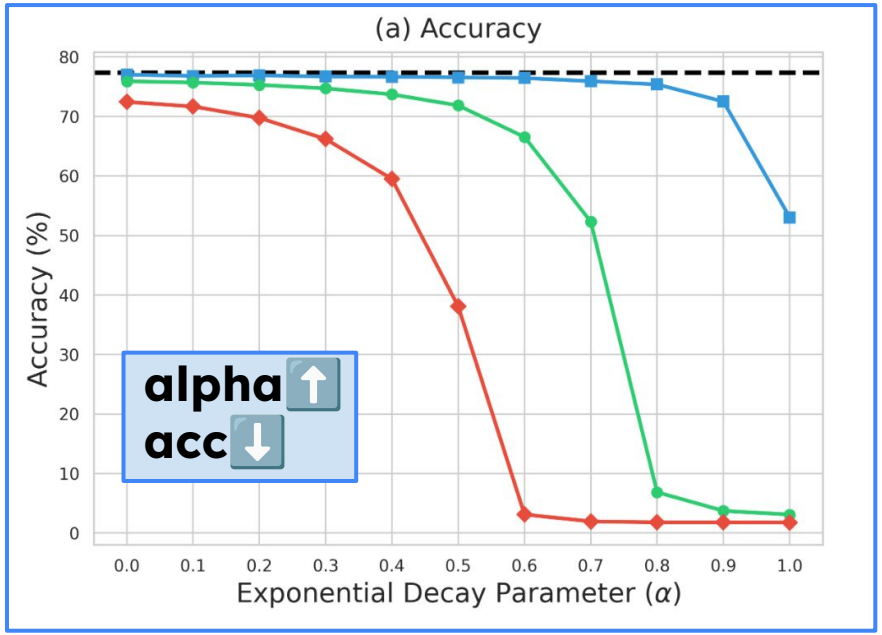
Model	Acc (%)	Params (M)	FLOPs (G)	Throughput (img/s) [GPU | BS=128]	Peak Memory (MB) [GPU | BS=128]
SReT-Tiny-Distill [4]	77.42	4.76	1.91	1072.86	795.76
SReT-Tiny-Distill+ToMe **(r_const=10)**	71.01 -6.41	-	1.32 -30.9%	1176.27 +9.6%	769.79 -3.3%
SReT-Tiny-Distill+ToMe **(r_lin=10)**	74.64 -2.78	-	1.46 -23.6%	1177.32 +9.7%	756.22 -5.0%

Results - Exponential Reduction

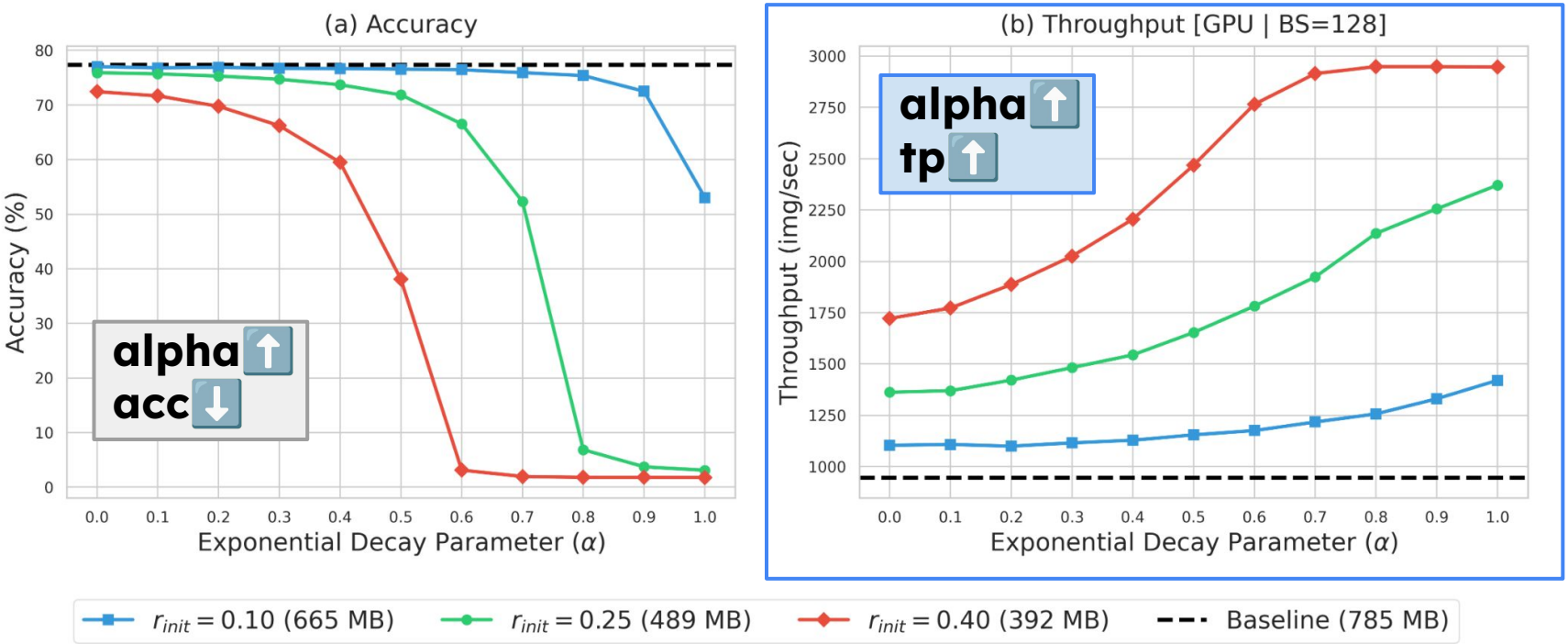
Merge an initial fraction of tokens **N x (exponential coefficient)** at first layer, then decay that amount by factor of **(alpha)** at each subsequent layer

Model	Acc (%)	Params (M)	FLOPs (G)	Throughput (img/s) [GPU | BS=128]	Peak Memory (MB) [GPU | BS=128]
SReT-Tiny-Distill [4]	77.42	4.76	1.91	1072.86	795.76
SReT-Tiny-Distill+ToMe **(r_const=10)**	71.01 -6.41	-	1.32 -30.9%	1176.27 +9.6%	769.79 -3.3%
SReT-Tiny-Distill+ToMe **(r_lin=10)**	74.64 -2.78	-	1.46 -23.6%	1177.32 +9.7%	756.22 -5.0%
SReT-Tiny-Distill+ToMe **(r_init=0.25, alpha=0)**	75.95 -1.47	-	1.49 -22.0%	1368.67 +27.6%	489.48 -38.5%

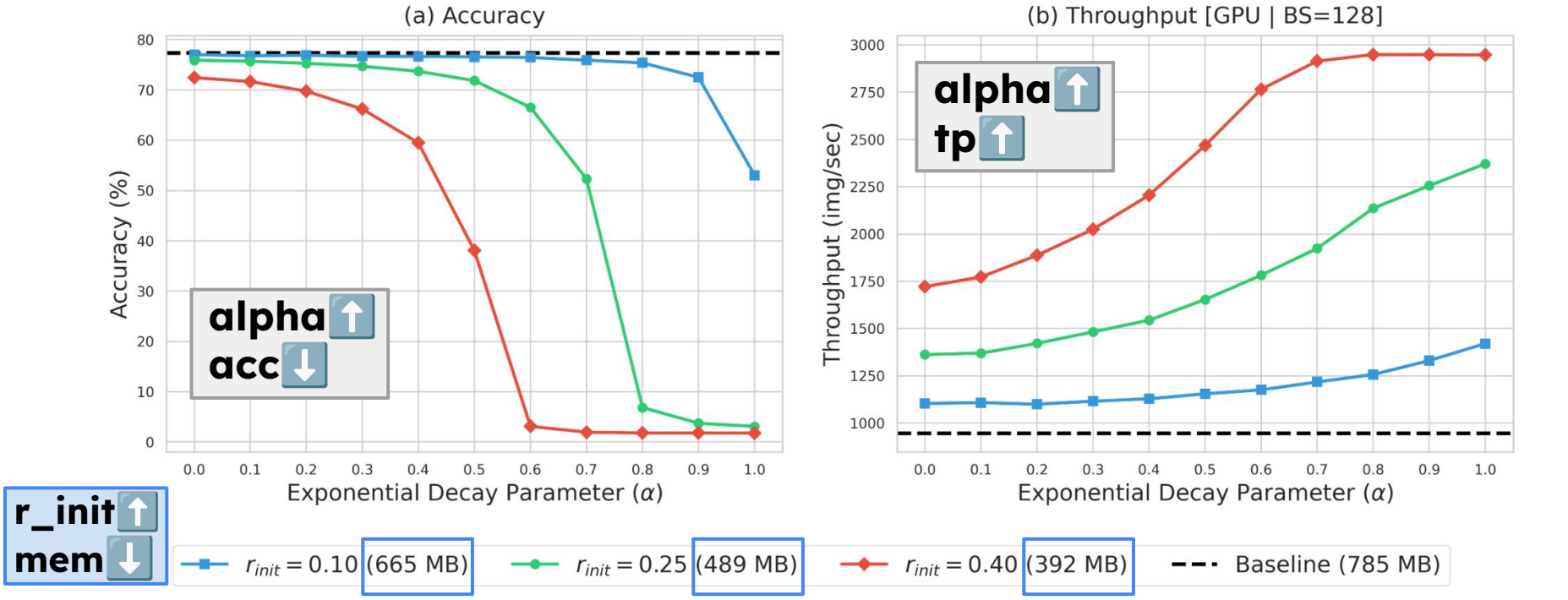
Evaluation - Impact of Decay Parameter (alpha)



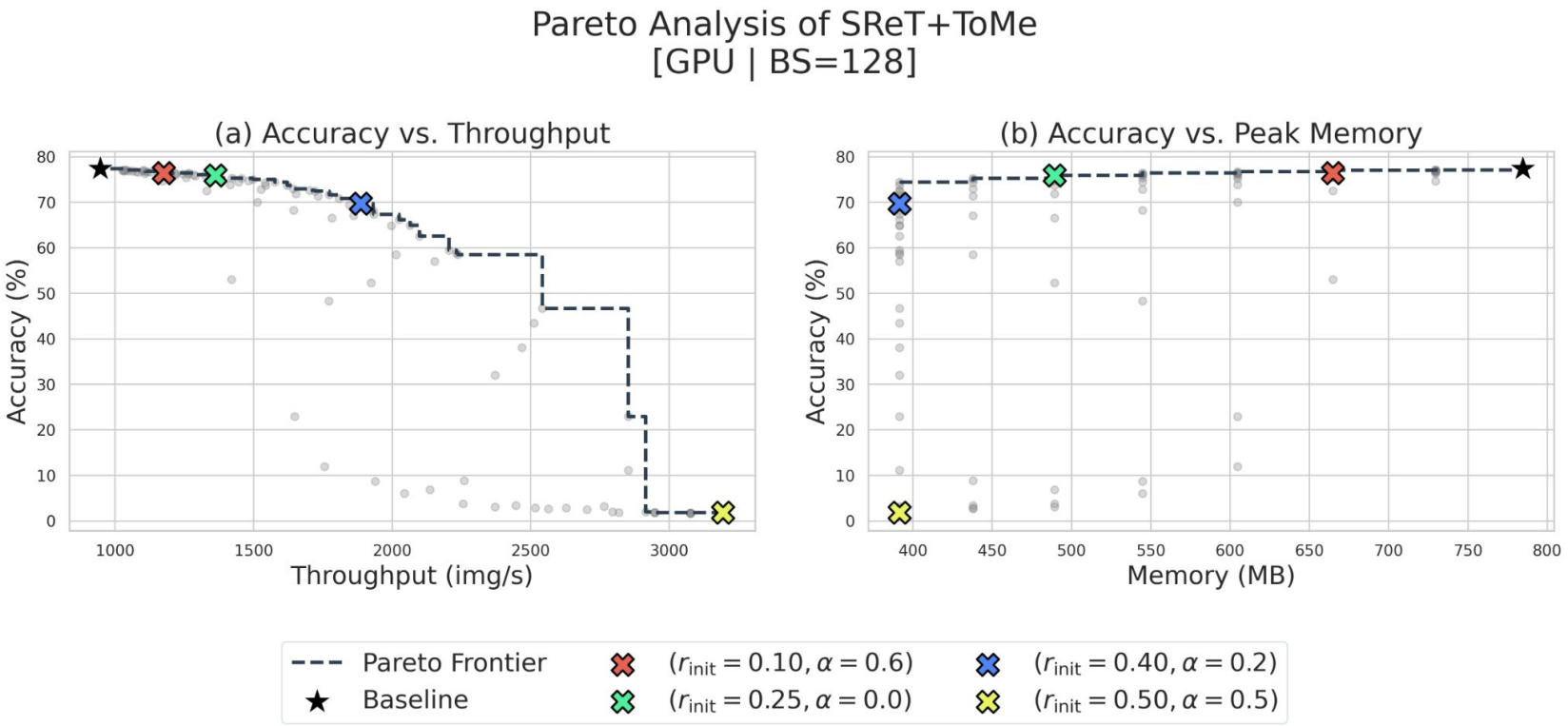
Evaluation - Impact of Decay Parameter (alpha)



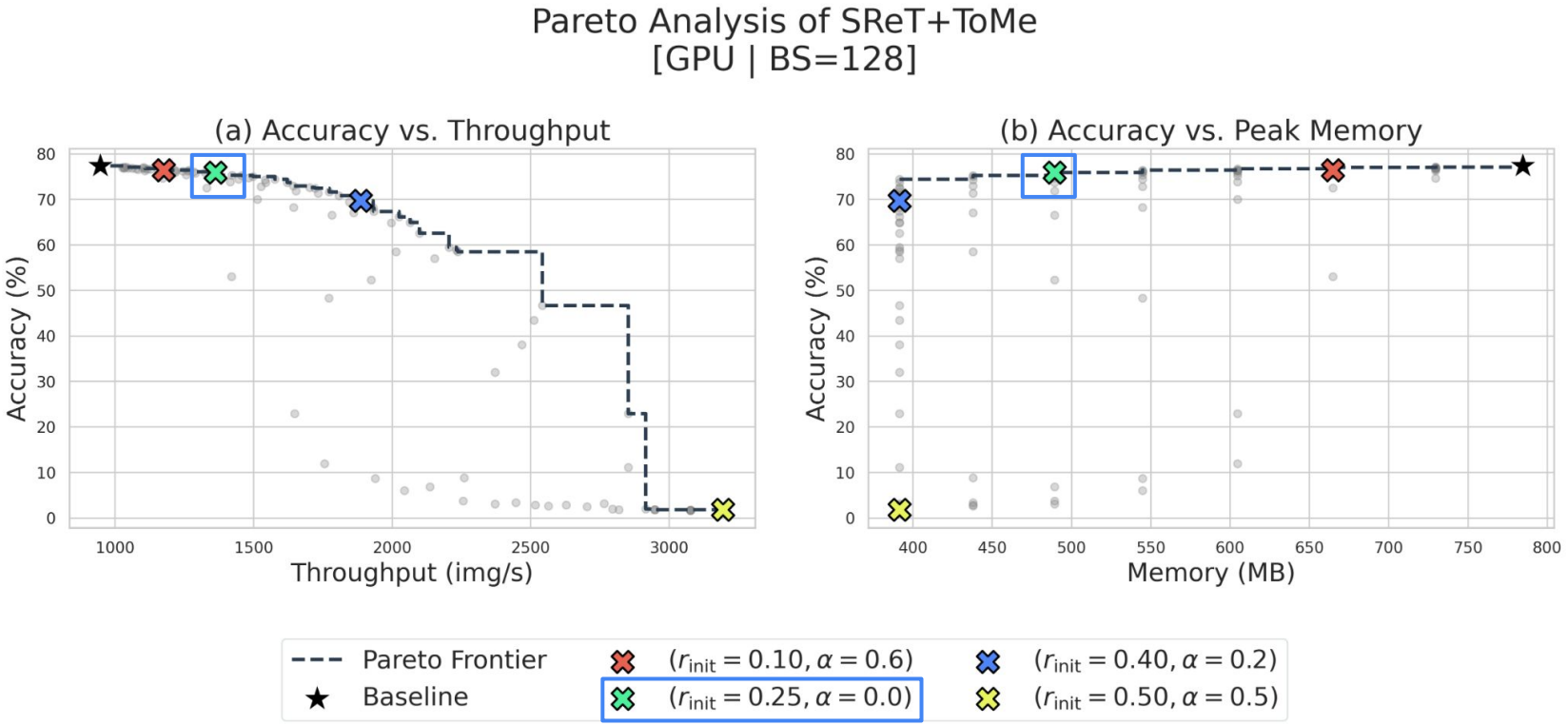
Evaluation - Impact of Decay Parameter (alpha)



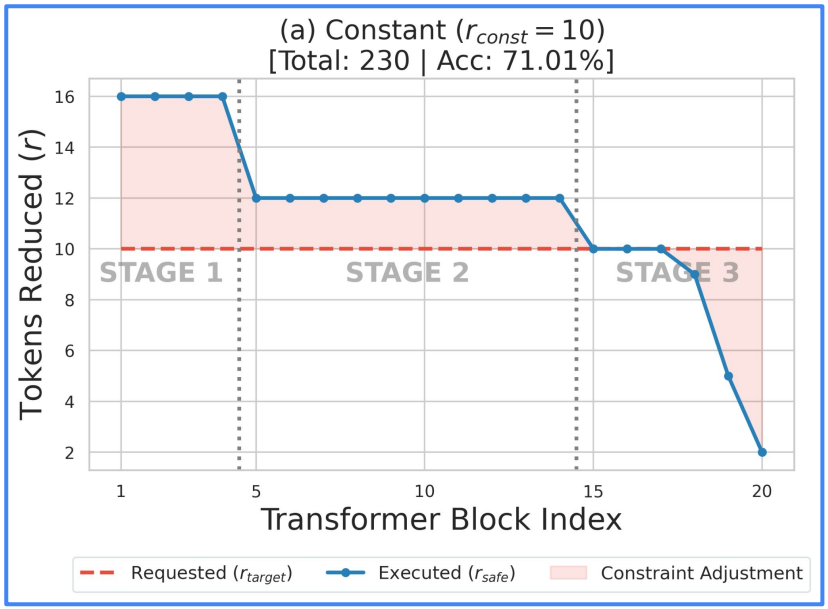
Evaluation - Pareto Analysis



Evaluation - Pareto Analysis



Evaluation - Reduction Schedule Performance

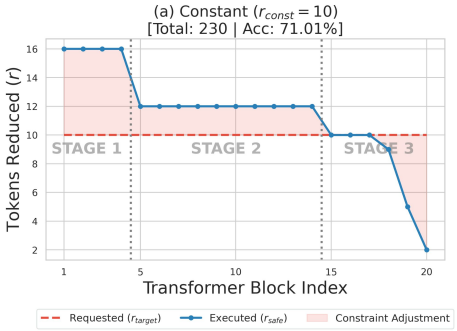


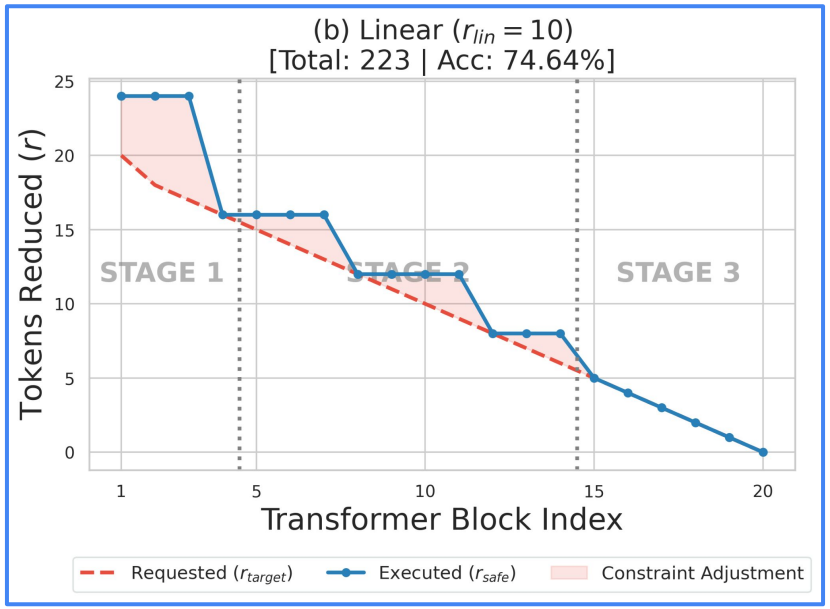
Constant (r=10)

Total tokens removed: 230

Accuracy: 71.01%

Evaluation - Reduction Schedule Performance





Constant (r=10)

Total tokens removed: 230

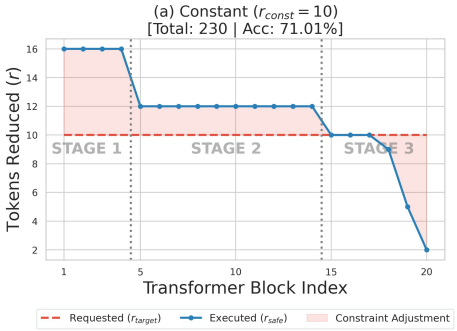
Accuracy: 71.01%

Linear (r=10)

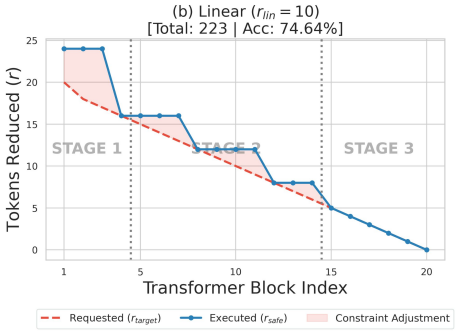
Total tokens removed: 223

Accuracy: 74.64%

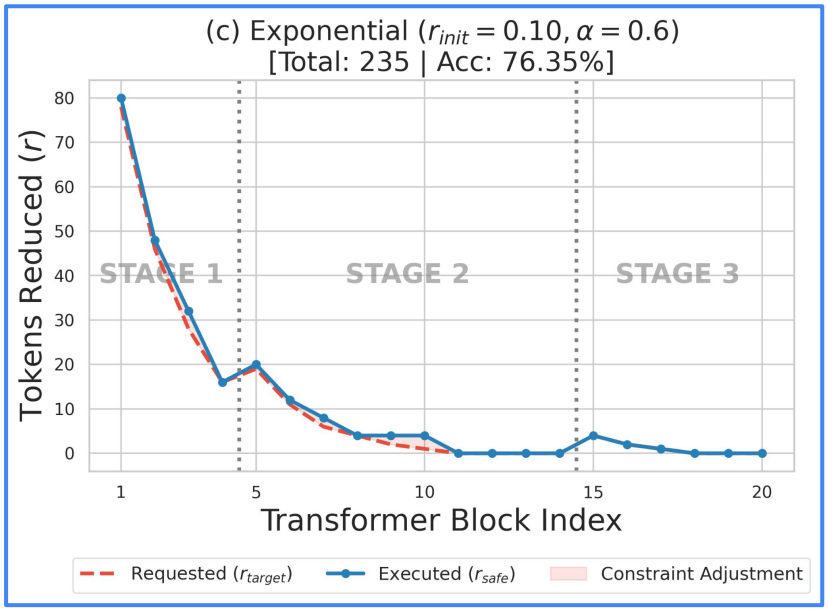
Evaluation - Reduction Schedule Performance



Constant (r=10)
Total tokens removed: 230
Accuracy: 71.01%



Linear (r=10)
Total tokens removed: 223
Accuracy: 74.64%



Exponential (r=0.1, a=0.6)
Total tokens removed: 235
Accuracy: 76.35%

Conclusions

✅SReT + ToMe Integration

- Training-free

✅Exponential Reduction Schedule

Model	Acc (%)	Params (M)	FLOPs (G)	Throughput (img/s) [GPU | BS=128]	Peak Memory (MB) [GPU | BS=128]
SReT-Tiny-Distill [4]	77.42	4.76	1.91	1072.86	795.76
SReT-Tiny-Distill+ToMe **(r_init=0.25, alpha=0)**	75.95 -1.47	-	1.49 -22.0%	1368.67 +27.6%	489.48 -38.5%

Conclusions

✅SReT + ToMe Integration

- Training-free

✅Exponential Reduction Schedule

⚠️**Used a Specific Model (SReT)**

- **Other recursive models?**

⚠️**Lacks CPU/Edge Device Testing**

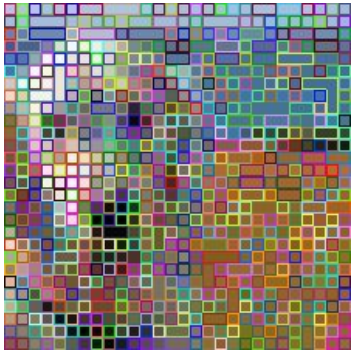
- **Algorithmic overhead?**

Model	Acc (%)	Params (M)	FLOPs (G)	Throughput (img/s) [GPU | BS=128]	Peak Memory (MB) [GPU | BS=128]
SReT-Tiny-Distill [4]	77.42	4.76	1.91	1072.86	795.76
SReT-Tiny-Distill+ToMe **(r_init=0.25, alpha=0)**	75.95 -1.47	-	1.49 -22.0%	1368.67 +27.6%	489.48 -38.5%

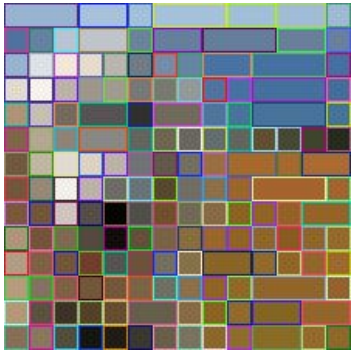
Thank you! Questions?

Input

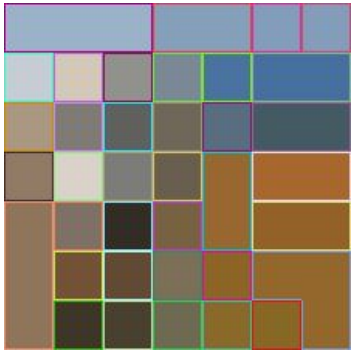
Stage 1



Stage 2



Stage 3



Results

Model	Acc (%)	Params (M)	FLOPs (G)	Throughput (img/s) [GPU | BS=128]	Peak Memory (MB) [GPU | BS=128]
SReT-Tiny-Distill [4]	77.42	4.76	1.91	1072.86	795.76
SReT-Tiny-Distill+ToMe **(r_const=10)**	71.01 -6.41	-	1.32 -30.9%	1176.27 +9.6%	769.79 -3.3%
SReT-Tiny-Distill+ToMe **(r_const=20)**	41.06 -36.36	-	1.06 -44.5%	1331.23 +24.1%	756.22 -5.0%
SReT-Tiny-Distill+ToMe **(r_lin=10)**	74.64 -2.78	-	1.46 -23.6%	1177.32 +9.7%	756.22 -5.0%
SReT-Tiny-Distill+ToMe **(r_lin=20)**	14.87 -62.55	-	1.07 -44.0%	1427.14 +33.0%	729.46 -8.3%
SReT-Tiny-Distill+ToMe **(r_init=0.10, α=0.6)**	76.35 -1.07	-	1.61 -15.7%	1186.62 +10.6%	664.75 -16.5%
SReT-Tiny-Distill+ToMe **(r_init=0.25 α=0.0)**	75.95 -1.47	-	1.49 -22.0%	1368.67 +27.6%	489.38 -38.5%
SReT-Tiny-Distill+ToMe **(r_init=0.40, α=0.2)**	69.71 -7.71	-	1.13 -40.8%	1886.90 +75.9%	391.51 -50.8%

References

[1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token Merging: Your ViT But Faster. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2210.09461

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[3] 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. In International Conference on Learning Representations. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[4] Zhiqiang Shen, Zechun Liu, and Eric Xing. 2022. Sliced Recursive Transformer. In Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV (Tel Aviv, Israel). Springer-Verlag, Berlin, Heidelberg, 727–744. https://doi.org/10.1007/978-3-031-20053-3_42

[5] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning. PMLR, 10347–10357. https://proceedings.mlr.press/v139/touvron21a.html