Semantic Equitable Clustering: A Simple and Effective Strategy for Clustering Vision Tokens

Supplementary Material

A. Experimental Details

SECViT's Architectures. SECViT's architecture details are illustrated in Table 1. In SECViT, we adopt four 3×3 convolutions to embed the input image into tokens, batch normalization and GELU are used after each convolution. 3×3 convolutions with stride 2 are used between stages to reduce the feature resolution. 3×3 DWConvs are adopted in CPE. For all models, we set the number of clusters in the first three stages to 32, 8, and 2, respectively.

B. More experimental Results

Efficiency Comparison. In Tab. 2, we compare the inference efficiency of various models in detail. From this, we can see that the ViT based on SEC demonstrates the best performance-speed tradeoff.

Different Methods for Merging Vision Tokens. For MLLM, SEC uses an interleaved merge token approach to reduce the number of vision tokens. Conversely, we also explore a sequential merge token method to achieve a similar reduction. The comparison of these two methods is shown in Tab. 3. The direct sequential merge token approach may result in the loss of critical visual information, significantly degrading the model's performance.

C. More Clustering Results for Complex Scenes

To further illustrate the mechanism of SEC, we visualize more images in complex scenes and their clustering results, as shown in Fig. 1. Specifically, we visualize the clustering results of the first three stages of SECViT. The results further demonstrate that SEC can better learn fine-grained representations in the shallow layers of the model and semantic representations in the deeper layers.

Model	Blocks	Channels	Heads	Ratios	Params(M)	FLOPs(G)
SECViT-T	[2, 2, 9, 2]	[64, 128, 256, 512]	[2, 4, 8, 16]	3	15	2.5
SECViT-S	[4, 4, 18, 4]	[64, 128, 256, 512]	[2, 4, 8, 16]	3	27	4.6
SECViT-B	[4, 8, 26, 9]	[80, 160, 320, 512]	[2, 4, 8, 16]	3	57	9.8
SECViT-L	[4, 8, 26, 9]	[112, 224, 448, 640]	[4, 8, 14, 20]	3	101	18.2
SECViT-XL	[6, 12, 28, 12]	[128, 256, 512, 1024]	[4, 8, 16, 32]	3	205	36.4

Table 1. Detailed Architectures of our models.

Model	Params(M)	FLOPs(G)	Throughput(imgs/s)	Top1-Acc(%)
DeiT-S [11]	22	4.6	3204	79.8
EViT-DeiT-S (keeprate=0.9) [7]	22	4.0	3428	79.8
SEC-DeiT-S (num_cluster=4)	22	4.1	3412	80.5
DeiT-B [11]	86	17.6	1502	81.8
SEC-DeiT-B	86	14.8	1682	82.4
PVTv2-b1 [13]	13	2.1	2204	78.7
TCFormer-light [16]	14	3.8	417	79.4
MPViT-XS [6]	11	2.9	1496	80.9
BiFormer-T [17]	13	2.2	1634	81.4
CMT-XS [4]	15	1.5	1476	81.8
GC-ViT-XT [5]	20	2.6	1308	82.0
SMT-T [8]	12	2.4	638	82.2
RMT-T [2]	14	2.5	1438	82.4
SECViT-T	15	2.5	2004	82.7
Swin-T [9]	29	4.5	1723	81.3
PS-ViT-B14 [15]	21	5.4	1986	81.7
DVT-T2T-ViT-19 [14]	39	6.2	1268	81.9
SGFormer-S [3]	23	4.8	952	83.2
CMT-S [4]	25	4.0	846	83.5
CSwin-S [1]	35	6.9	972	83.6
SMT-S [8]	20	4.8	356	83.7
BiFormer-S [17]	26	4.5	766	83.8
SEC-Swin-T	29	4.8	1482	83.8
SECViT-S	27	4.6	998	84.3
Swin-S [9]	50	8.8	1006	83.0
SGFormer-M [3]	39	7.5	598	84.1
SMT-B [8]	32	7.7	237	84.3
BiFormer-B [17]	57	9.8	498	84.3
MaxViT-S [12]	69	11.7	546	84.5
CMT-B [4]	46	9.3	447	84.5
iFormer-B [10]	48	9.4	688	84.6
RMT-B [2]	54	9.7	430	85.0
SEC-Swin-S	50	9.2	804	85.0
SECViT-B	57	9.8	504	85.2
Swin-B [9]	88	15.5	768	83.5
CSWin-B [1]	78	15.0	660	84.2
SMT-L [8]	80	17.7	158	84.6
SGFormer-B [3]	78	15.6	388	84.7
iFormer-L [10]	87	14.0	410	84.8
MaxViT-B [12]	120	23.4	306	84.9
SEC-Swin-B	88	16.2	696	85.3
SECViT-L	101	18.2	398	85.7

Table 2. Comparison of models' efficiency. Throughputs are measured on a single A100 with the batch size of 64.

Method	V-T num	Time	Speed	TextVQA	GQA	VQAv2	POPE	MM-Vet
Interleaved	288+1	14h	1.5×	60.1	63.5	78.9	87.7	33.2
Sequential	288+1	14h	1.5×	52.8(-7.3)	57.1(-6.2)	75.7(-3.2)	81.7(-6.0)	27.6(-5.6)
Interleaved	144+1	10h	2.1×	56.8	62.0	78.0	86.1	31.7
Sequential	144+1	10h	2.1×	47.2(-9.6)	53.6(-8.4)	71.7(-6.3)	80.0(-6.1)	22.3(-9.6)

Table 3. Different methods for merging vision tokens.

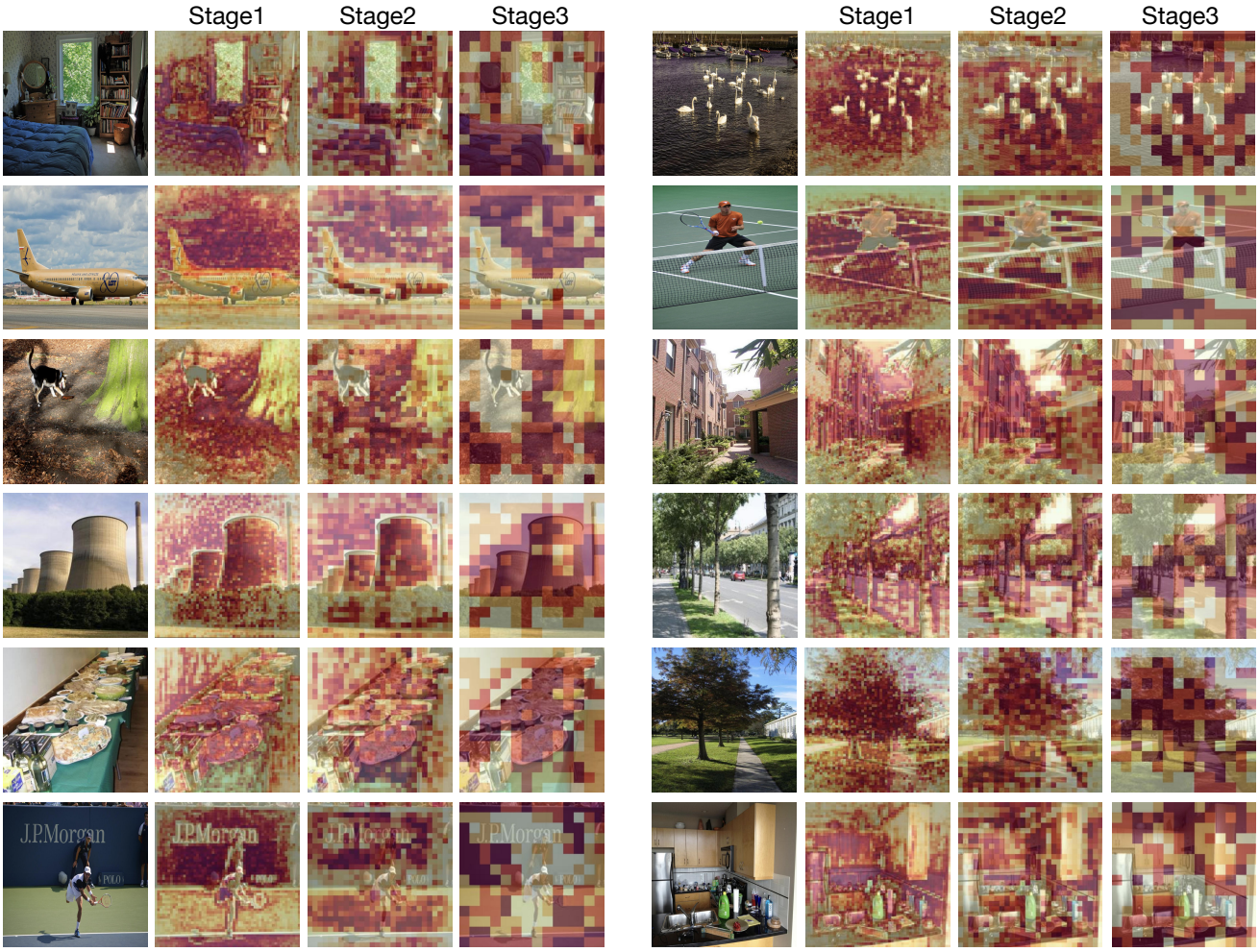

Figure 1. Visualization results for complex scenes.

References

[1] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, 2022. 2

[2] Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, and Ran He. Rmt: Retentive networks meet vision transformers. In *CVPR*, 2024. 2

[3] SG-Former: Self guided Transformer with Evolving Token Reallocation. Sucheng ren, xingyi yang, songhua liu, xinchao wang. In *ICCV*, 2023. 2

[4] Jianyuan Guo, Kai Han, Han Wu, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Cmt: Convolutional neural networks meet vision transformers. In *CVPR*, 2022. 2

[5] Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In *ICML*, 2023. 2

[6] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *CVPR*, 2022. 2

[7] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022. 2

[8] Weifeng Lin, Ziheng Wu, Jiayu Chen, Jun Huang, and Lianwen Jin. Scale-aware modulation meet transformer. In *ICCV*, 2023. 2

[9] 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. 2

[10] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng YAN. Inception transformer. In *NeurIPS*, 2022. 2

[11] Hugo Touvron, Matthieu Cord, Matthijs Douze, et al. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 2

[12] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV*, 2022. 2

[13] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):1–10, 2022. 2

[14] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic vision transformers with adaptive sequence length. In *NeurIPS*, 2021. 2

[15] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip HS Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. In *ICCV*, 2021. 2

[16] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *CVPR*, 2022. 2

[17] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson Lau. Biformer: Vision transformer with bi-level routing attention. In *CVPR*, 2023. 2