6. Supplementary

Sharing FFN when grafting: Table 9 compares the trade-offs of having either a shared FFN or two separate FFNs, using a DeiT-T backbone. When having separate FFNs, GrafT features are fused after FFNs, and when having a shared FFN, fusion happens before it. We see that a shared FFN achieves +1.3% higher accuracy with fewer parameters compared to its counterpart. Therefore, we adopt a shared FFN design in the GrafT by default.

Table 9. Sharing the parameters of backbone-FFN by training the DeiT-T on ImageNet-1K

Shared FFN	Params (M) ↓	FLOPs (G) ↓	Acc. (%) ↑
✗	8.2	1.2	74.8
✓	7.9	1.2	**(+1.3) 76.1**

Relative performance of GrafT with various Transformers: Table 10 shows the backbones in which we integrate GrafT and their architectural characteristics. Backbones cover the combination of two types of vertical structures (Pyramid/Homogeneous), two types of Transformers (Hybrid/Pure), and five types of self-attention mechanisms.

Table 10. Characteristics of backbones in terms of vertical structures (Homogeneous/Pyramid), Transformer type (Hybrid/Pure Transformer), and attention mechanisms.

Model	Ver. Struc.	Type	Attn. method
MViTv2 [22]	Pyramid	Hybrid	MHSA
MobViT [29]	Pyramid	Hybrid	Inter-patch
MobViTv2 [30]	Pyramid	Hybrid	Separable
Swin [27]	Pyramid	Transformer	Shifted window
CSWin [9]	Pyramid	Transformer	Cross-shaped
DeiT [37]	Homogeneous	Transformer	MHSA

Relative performance of GrafT with mobile backbones on object detection: Table 11 shows the relative performance of GrafT with mobile Transformers on a single shot object detection task. GrafT improves the mAP of MobViT by (+0.7% for -XXS), (+1.6% for -XS), (+1.1% for -S) with the small addition of complexities. The correponding increase in (parameters, FLOPs) pairs are (+9%, 6%), (+12%, 5%), (+14%, 5%), respectively. GrafT boosts the mAP of MobViTv2-0.5 by +1.7% while incurring +8% more parameters and 2% more FLOPs. It shows that GrafT is a light-weight module supporting mobile Transformers to become general-purpose backbones.

Table 11. Relative performance of GrafT with mobile backbones on a single shot object detection task on the COCO 2017 [25]. GrafT consistently improves the detection performance of MobViT [29] and MobViTv2 [30].

Model	Type	Params ↓ (M)	FLOPs ↓ (G)	Acc. ↑ (%)
MobViT-XXS [29]	Hybrid	1.7	0.90	19.9
MobViT-XXS+GrafT	Hybrid	1.9	0.91	(+0.7) 20.6
MobViTv2-0.5	Hybrid	2.0	0.92	19.9
MobViTv2-0.5+GrafT	Hybrid	2.2	0.94	(+1.7) 21.6
MobileViT-XS [29]	Hybrid	2.7	1.89	24.8
MobileViT-XS+GrafT	Hybrid	3.1	1.98	(+1.6) 26.4
MobileViT-S [29]	Hybrid	5.7	3.48	27.7
MobileViT-S+GrafT	Hybrid	6.5	3.65	(+1.1) 28.8

Visualization of self-attention scores: In Figure 4, we visualize the self-attention maps to understand the benefit of integrating GrafT. Layer 2, Layer 5, and Layer 10 within a 12-layer model are used to analyze the self-attention maps at shallow, middle, and deep layers in Transformers. We evaluate DeiT-T+GrafT on the validation images in ImageNet-1K to draw self-attention scores. In the first row of Figure 4, self-attention captures the overall shape of cows, human, fences, and trees while being a bit inaccurate in highlights at Layer 2, focuses on cows and human with a more accuracy at Layer 5, and attends to only important parts of cows at Layer 10. In the second row of Figure 4, self-attention captures the overall shape of mice focusing on outlines of mice with a bit inaccurate highlights at Layer 2, and refines the outlines at Layer 5 and Layer 10. It shows that GrafT can provide multi-scale high-level semantics (i.e., capture global features) even within shallow layers.

Comparison of multi-scale tokens: Table 12 summarizes the difference between GrafT and previous works on how to deliver high-level semantics. In the homogeneous structure, GrafT adopts average pooling as downsampling and learnable bilinear interpolation as upsampling. It is a faster mechanism than cross attention in CrossViT and it provides the flexibility of creating various sizes of feature maps. In the pyramid structure, GrafT is unique in the sense that it creates multi-scale features at each layer whose grid sizes are the same as vertical multi-scale features. For example, Swin-T+GrafT exploits four different scales of features in each layer at the first stage, as there are four vertical stages. On the other hand, other models exploit at most two scales of features. The fusion mechanism of horizontal multi-scale features follows the consecutive element-wise addition in FPN [24].

Figure 4. The visualization of the scores of attention maps at different layers of DeiT-T+GrafT. The input images are from the validation set in ImageNet-1K. GrafT provides multi-scale high-level semantics to the backbone to capture the global features from the early-stage layer.

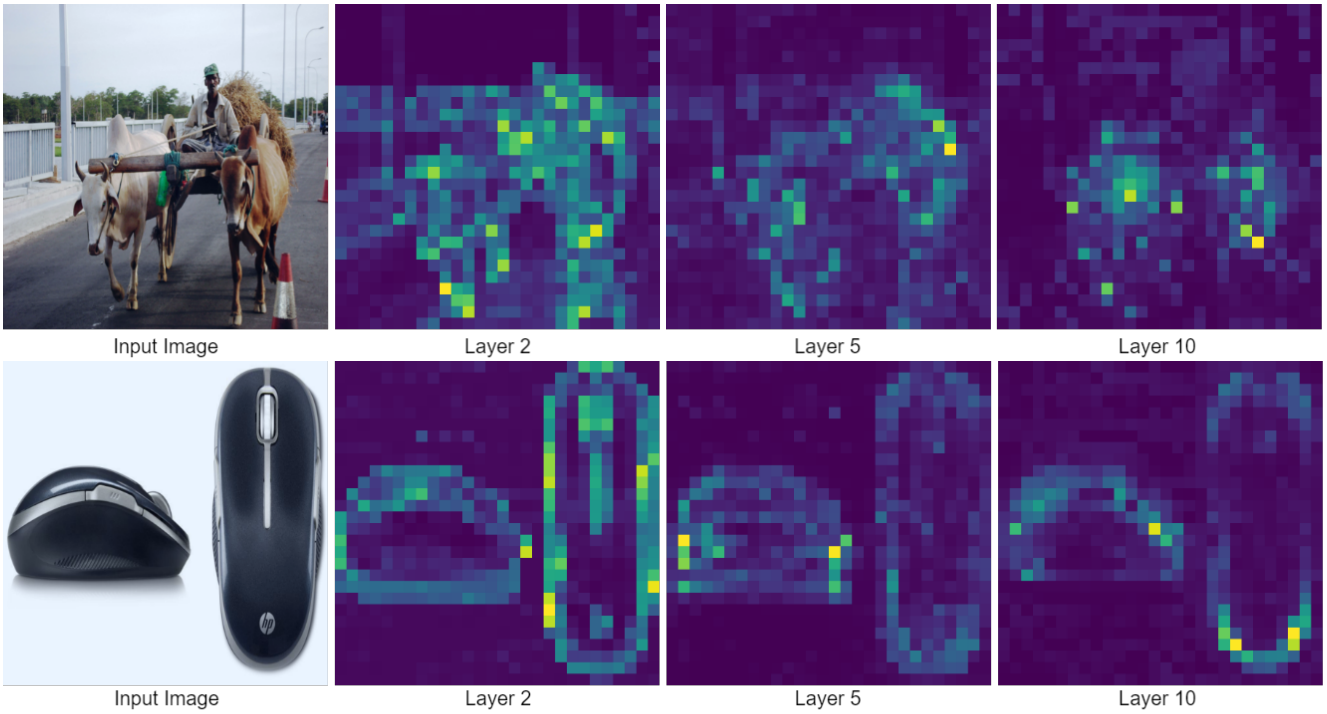

Table 12. Approaches to deliver high-level semantics in terms of their vertical structure, #scales per layer and fusion method.

Model	Vertical structure	#scales	Fusion method
ViT-T (DeiT-T) [37]	Homogeneous	None	None
CrossViT-9 [4]	Homogeneous	2	Cross attention
DeiT-T + GrafT	Homogeneous	2	Learn. W-Bilinear + E-Wise add.
PVT-T [38]	Pyramid	None	None
T2T$_t$ [42]	Homogeneous	None	None
PoolFormer-S12 [41]	Pyramid	None	None
TNT-S [12]	Homogeneous	2	LL + E-Wise add.
Swin-T [27]	Pyramid	None	None
RegionViT-S [3]	Pyramid	2	Cross attention
Swin-T + GrafT	Pyramid	4	Learn. W-Bilinear + E-Wise add.

Upsampling components: Table 13 presents the effect of discrete elements within the upsampling mechanism as integrated into the GrafT framework. Through the incorporation of both channel mixing and anti-aliasing components, the GrafT model attains 76.1% accuracy with 7.9M parameters and 1.2G FLOPs. In cases where the anti-aliasing or channel mixing elements are individually omitted, the reduction in accuracy by 0.4% or 1.0%, respectively, is observed with a marginal decrease in parameters. It underscores that both anti-aliasing and channel mixing are effective when employed in conjunction with bilinear interpolation within the GrafT.

Table 13. The efficacy of incorporating channel mixing and anti-aliasing elements within the upsampling mechanism.

Channel mixing	Anti-aliasing	Params (M) ↓	FLOPs (G) ↓	Acc. (%) ↑
✓	✓	7.9	1.2	76.1
✓	✗	7.8	1.2	(-0.4)75.7
✗	✓	7.5	1.2	(-1.0)75.1

Table 14. Replacing local self-attention by convolution module in GrafT. The GrafT with local self-attention achieves 1.1% better accuracy with fewer FLOPs and a marginal increase in parameters

Model	Params (M)↓	FLOPs (G)↓	Acc. (%)↑
DeiT-T+GrafT	7.9	1.2	76.1
DeiT-T+IR	7.4	1.5	75.0
DeiT-T	5.7	1.3	72.2

Replacing GrafT with convolution modules: Table 14 shows the effectiveness of the current GrafT design compared to convolution modules. The inverted residual module (IR) from MobileNetv2 are attached to the DeiT-T backbone instead of GrafT and trained on ImageNet-1K. The current GrafT design outperforms DeiT-T+IR by 1.1% with smaller FLOPs and a marginal increase in Parameters.

Training details for image classification: ImageNet-1K [8] is a classification benchmark with annotations of 1000 categories. It contains 1.2M training images and 50K validation images. In our evaluation, we report Top-1 accuracy (%) on a single-crop setting along with complexity metrics (measured in Parameters and FLOPs). We train our models with the standard settings. For pure Transformers [9, 27, 37], we run 300 epochs with 224×224 resolution inputs, using `timm` [39] library. For hybrid Transformers [29,30], we run 300 epochs with a multi-scale sampler ranging from 160 to 320 inputs with step-size 32, using `CVNets` [28] library. We consider the original hyperparameter settings for each of the backbones and apply stochastic depths. In MobViT, stochastic depths are 0.0, 0.1, 0.2 for -XXS, -XS, -S. In MobViTv2, stochastic depths are 0.0 for both v2-0.5, v2-1.0. In DeiT-T, stochastic depth is 0.0. In Swin, stochastic depths are 0.25, 0.4 for -T, -S. In CSWin, stochastic depths are 0.1, 0.35 for -XT and -T. Please note that we design CSWin-XT* by modifying CSWin-T by reducing channels to (48, 96, 192, 384) and setting the number of layers to (1, 2, 7, 1) in the four stages.

Training details for semantic segmentation: ADE20K [44] annotates 150 categories for semantic segmentation. It contains 20K training, 2K validation and 3K testing images. In our evaluations, we use multi-scale mIoU as the metric (using scales of [0.5, 0.75, 1.0, 1.25, 1.5, 1.75]$\times$ the training resolution) and follow a training procedure similar to Swin [27]. In Table 5, we also report model complexity metrics such as parameters, FLOPs (for an input size of 512×2048). We use GrafT backbones pre-trained on ImageNet-1K [8] for 300 epochs at a 224×224 resolution, and finetune it with the decoder at a 512×512 resolution. We choose UperNet [40] as our decoder, and implement within the `mmsegmentation` [7] framework. Swin-T+GrafT uses the stochastic depth of 0.2.

Training details for object detection: The COCO 2017 dataset [25] consists of 118K images for training, 5K for validation, and 20K for testing. In a single shot object detection, We use SSDLite [26, 29], a light-weight object detection backbone, and exploit (320x320) input to finetune MobViT+GrafT and MobViTv2+GrafT on the dataset. MobViT-XXS, MobViT-XS, MobViT-S, MobViTv2-0.5 use stochastic depths of 0.0, 0.1, 0.1, 0.0 respectively and follow standard settings as in MobViT [29] and MobViTv2 [30]. In two-stage object detection, we use Mask R-CNN [13] framework to adopt Swin-T+GrafT pretrained on the ImageNet-1K. For training, we use the stochastic depth with ratios of 0.1 and 0.2 for $1\times$ (SS) and $3\times$ (MS) schedules, respectively and follow the original hyperparameter settings as in Swin [27]. Here, $1\times$ (SS) corresponds to 12 epochs with single scale, $3\times$ (MS) corresponds to 36 epochs with multi-scale.

References

[1] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019. 1

[2] Andrew Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-Performance Large-Scale Image Recognition Without Normalization. *arXiv*, 2021. 1

[3] Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. Regionvit: Regional-to-local attention for vision transformers. In *International Conference on Learning Representations*, 2022. 1, 6, 8, 10

[4] Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 357–366, October 2021. 1, 2, 5, 8, 10

[5] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. *arXiv preprint arXiv:2108.05895*, 2021. 6, 8

[6] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. *arXiv preprint arXiv:1707.01629*, 2017. 8

[7] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020. 11

[8] 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. 2, 5, 6, 7, 8, 11

[9] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12124–12134, June 2022. 2, 5, 6, 8, 9, 11

[10] 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. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 1, 2, 3, 5, 8

[11] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12175–12185, 2022. 8

[12] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing XU, and Yunhe Wang. Transformer in transformer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15908–15919. Curran Associates, Inc., 2021. 1, 6, 8, 10

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 11

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 8

[15] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11936–11945, October 2021. 8

[16] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. 6, 8

[17] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[18] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1, 2, 3, 6, 8

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 8

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 8

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 8

[22] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 2, 6, 8, 9

[23] Yanyu Li, Geng Yuan, Yang Wen, Eric Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *arXiv preprint arXiv:2206.01191*, 2022. 8

[24] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 3, 4, 9

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6, 7, 9, 11

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 6, 11

[27] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

[28] Sachin Mehta, Farzad Abdolhosseini, and Mohammad Rastegari. Cvnets: High performance library for computer vision. *CoRR*, 2022. 11

[29] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2022. 2, 5, 6, 7, 8, 9, 11

[30] Sachin Mehta and Mohammad Rastegari. Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*, 2022. 2, 5, 6, 8, 9, 11

[31] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 1, 6, 8

[32] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *Advances in Neural Information Processing Systems*, 35:23495–23509, 2022. 8

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 8

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 8

[36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1, 8

[37] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine*

Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021. 1, 2, 5, 8, 9, 10, 11

[38] Wenhai 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, October 2021. 5, 6, 8, 10

[39] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 11

[40] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 11

[41] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10819–10829, June 2022. 6, 10

[42] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021. 6, 8, 10

[43] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2998–3008, October 2021. 6, 8

[44] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 3, 6, 7, 11