Appendix

1 Performance on Dense Prediction Tasks

We also extend `ToCom` to semantic segmentation tasks. Since semantic segmentation needs to perform predictions on all tokens, we merge tokens before the FFN layers, and unmerge them after FFNs. We conduct experiments on ADE20k [32] datasets using pre-trained DeiT-B_{384} as encoder and Mask Transformer of Segmenter [25] as decoder. We sample $r \in \{0, 1, 2, ..., 16\}$ in training and merge $32r$ tokens before FFNs. Since the image resolution of ADE20k is 512×512, we resize the image of ImageNet to 512×512 when training `ToCom`, and reduce the number of training epoch to 2. We provide results of single-scale evaluation in Table 1 with source $r = 0$ and target $r = 8, 12, 16$. `ToCom` also improves the performance of token compression on dense prediction tasks.

Table 1: Performance on ADE20k validation set. We keep source $r = 0$. We also report GFLOPs of encoder.

Setting	mIoU (SS)	GFLOPs
Target $r = 0$	48.7	106.2
Target $r = 8$	48.0	91.8
+ `ToCom`	**48.3**	91.8
Target $r = 12$	46.4	84.5
+ `ToCom`	**47.2**	84.5
Target $r = 16$	41.3	77.3
+ `ToCom`	**43.4**	77.3

2 Pseudocode of Training

See Algorithm 1.

3 Experimental Details

3.1 PET for VTAB-1k

On VTAB-1k benchmark, we use PET method instead of full fine-tuning to obtain off-the-shelf models, since previous work [11,13,14,21] has found that PET performs much better than full fine-tuning on VTAB-1k. Especially, [14] finds that Adaptformer [2] outperforms other PET methods, and thus we also adopt it in experiments on VTAB-1k. AdaptFormer uses bottleneck FFN with in-between ReLU activation as adapters. The weights of an adapter are $\boldsymbol{W}_{down} \in \mathbb{R}^{d \times 32}$

Algorithm 1 Training `ToCom`

Require: Pre-trained model $\widehat{\mathcal{M}}$, pre-training dataset $\mathcal{D}$
  Initialize $\mathcal{P} = \{\mathcal{P}_{0\to1}, ..., \mathcal{P}_{15\to16}\}$
  for $\boldsymbol{x} \in \mathcal{D}$ **do**
    Ramdomly sample m, n, *s.t.* $m \neq n$
    $\widehat{\mathcal{M}}_m \leftarrow$ Apply ToMe with $r = m$ to $\widehat{\mathcal{M}}$
    $\widehat{\mathcal{M}}_n \leftarrow$ Apply ToMe with $r = n$ to $\widehat{\mathcal{M}}$
    if $m < n$ **then**
      $\mathcal{L} = \mathcal{L}_{KD}\left(\widehat{\mathcal{M}}_m \oplus \left(\bigoplus_{i=m}^{n-1} \mathcal{P}_{i\to(i+1)}\right), \widehat{\mathcal{M}}_n\right)$
    else
      $\mathcal{L} = \mathcal{L}_{KD}\left(\widehat{\mathcal{M}}_m \ominus \left(\bigoplus_{i=n}^{m-1} \mathcal{P}_{i\to(i+1)}\right), \widehat{\mathcal{M}}_n\right)$
    end if
    Update $\mathcal{P}$ according to $\mathcal{L}$
  end for

and $\boldsymbol{W}_{up} \in \mathbb{R}^{32\times d}$. Adapters are inserted into networks as shortcuts of the FFN blocks, *i.e.*, given an input $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, the computation of FFNs becomes

$$\boldsymbol{X}' = \underbrace{\boldsymbol{X} + \texttt{FFN}(\texttt{LN}(\boldsymbol{X}))}_{\text{Frozen}} + \underbrace{s \cdot \texttt{ReLU}(\boldsymbol{X}\boldsymbol{W}_{down})\boldsymbol{W}_{up}}_{\text{Tuned}} \tag{1}$$

where s is a hyper-parameter searched from {0.01, 0.1, 1, 10, 100} and `LN` is layer normalization.

3.2 Datasets

See Table 4. For VTAB-1k, each dataset contains 800 samples for training and 200 for validation. Following previous work [11–13, 21, 31], after searching the hyper-parameters, we tune the pre-trained model with all the 1,000 training and validation samples and report results evaluated on test-set. For other datasets that do not have official validation set, we randomly split 10% training example for hyper-parameters searching.

3.3 Pre-Trained Backbones

See Table 2.

3.4 Code Implementation

We use *PyTorch* and *timm* to implement all experiments on 8×V100 GPUs.

3.5 Data Augmentation

ImageNet & CIFAR100 & Few-shot learning Following [26], for training samples, we use color-jitter, RandAugmentation, and repeated augmentation; for validation/test samples, we resize them to 256×256, crop them to 224×224 at the center, and then normalize them with ImageNet's mean and standard deviation.

Table 2: Pre-Trained backbones.

Model	Pre-Training Dataset	Pre-Trained Weights
DeiT-B/16 [26]	ImageNet-1K	checkpoint
DeiT-S/16 [26]	ImageNet-1K	checkpoint
ViT-B/16 (MAE) [8]	ImageNet-1K	checkpoint
DeiT-B/16_{384} [26]	ImageNet-1K	checkpoint

Table 3: Hyper-parameters.

Methods	ImageNet	CIFAR	FGVC	VTAB-1k
Epochs	10	100	30	100
Batch size	1024	1024	64	64
Optimizer	AdamW	AdamW	AdamW	AdamW
Base learning rate	1e-3	5e-5	{5e-3,2e-3,1e-3,5e-4,2e-4,1e-4}	-
Learning rate	-	-	-	1e-3
Learning rate decay	cosine	cosine	cosine	cosine
Weight decay	0.05	0.05	0.05	0.0001
Warmup epochs	0	10	5	0
Label smoothing ε	-	0.1	0.1	-
Stoch. Depth	0.1	0.1	0.1	0.1
Rand Augment	9/0.5	9/0.5	9/0.5	-
Mixup prob.	0.8	0.8	0.8	-
Cutmix prob.	1.0	1.0	1.0	-
Erasing prob.	0.25	0.25	0.25	-

VTAB-1K Following [11], we just resize the images to 224×224.

Semantic Segmentation We completely follow the setting used in [25], which does mean substraction, random resizing, random left-right flipping, and randomly crops large images and pad small images to 512×512.

3.6 Hyper-parameters

See Table 3.

Table 4: Statistics of used datasets.

	Dataset	# Classes	Train	Val	Test
	VTAB-1K [30]				
Natural	CIFAR100 [19]	100	800/1,000	200	10,000
	Caltech101 [6]	102			6,084
	DTD [4]	47			1,880
	Oxford-Flowers102 [23]	102			6,149
	Oxford-Pets [24]	37			3,669
	SVHN [22]	10			26,032
	Sun397 [29]	397			21,750
Specialized	Patch Camelyon [27]	2	800/1,000	200	32,768
	EuroSAT [9]	10			5,400
	Resisc45 [3]	45			6,300
	Retinopathy [16]	5			42,670
Structured	Clevr/count [15]	8	800/1,000	200	15,000
	Clevr/distance [15]	6			15,000
	DMLab [1]	6			22,735
	KITTI-Dist [7]	4			711
	dSprites/location [10]	16			73,728
	dSprites/orientation [10]	16			73,728
	SmallNORB/azimuth [20]	18			12,150
	SmallNORB/elevation [20]	18			12,150
	FGVC				
	CUB-200-2011 [28]	200	5,994	-	5,794
	Stanford Cars [18]	196	8,144	-	8,041
	Oxford Flowers102 [23]	102	1,020	1,020	6,149
	Stanford Dogs [17]	129	12,000	-	8,580
	Others				
	CIFAR100 [19]	100	60,000	-	10,000
	ImageNet [5]	1,000	1,281,167	50,000	-
	ADE20k [32]	150	20,210	2,000	3,352

References

1. Beattie, C., Leibo, J.Z., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., Lefrancq, A., Green, S., Valdés, V., Sadik, A., et al.: Deepmind lab. arXiv preprint **arXiv:1612.03801** (2016) 4
2. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. In: Proceedings of NeurIPS (2022) 1
3. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proc. IEEE (2017) 4
4. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , Vedaldi, A.: Describing textures in the wild. In: Proceedings of CVPR (2014) 4
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of CVPR (2009) 4
6. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: Proceedings of CVPR workshops (2004) 4
7. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research (2013) 4
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: Proceedings of CVPR (2022) 3
9. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2019) 4
10. Higgins, I., Matthey, L., Pal, A., Burgess, C.P., Glorot, X., Botvinick, M.M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: Proceedings of ICLR (2017) 4
11. Jia, M., Tang, L., Chen, B., Cardie, C., Belongie, S.J., Hariharan, B., Lim, S.: Visual prompt tuning. In: Proceedings of ECCV (2022) 1, 2, 3
12. Jie, S., Deng, Z.: Convolutional bypasses are better vision transformer adapters. arXiv preprint **arXiv:2207.07039** (2022) 2
13. Jie, S., Deng, Z.H.: Fact: Factor-tuning for lightweight adaptation on vision transformer. In: Proceedings of AAAI (2023) 1, 2
14. Jie, S., Wang, H., Deng, Z.: Revisiting the parameter efficiency of adapters from the perspective of precision redundancy. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 17171–17180. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.01579, https://doi.org/10.1109/ICCV51070.2023.01579 1
15. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of CVPR (2017) 4
16. Kaggle, EyePacs: Kaggle diabetic retinopathy detection (2015), https://www.kaggle.com/c/diabetic-retinopathy-detection/data 4
17. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization: Stanford dogs (2012) 4
18. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of CVPR workshops (2013) 4
19. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 4

20. LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of CVPR (2004) 4
21. Lian, D., Zhou, D., Feng, J., Wang, X.: Scaling & shifting your features: A new baseline for efficient model tuning. In: Proceedings of NeurIPS (2022) 1, 2
22. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: Proceedings of NIPS Workshops (2011) 4
23. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: Proceedings of CVPR (2006) 4
24. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: Proceedings of CVPR (2012) 4
25. Strudel, R., Pinel, R.G., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of ICCV (2021) 1, 3
26. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Proceedings of ICML (2021) 2, 3
27. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology. arXiv preprint **arXiv:1806.03962** (2018) 4
28. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011) 4
29. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Proceedings of CVPR (2010) 4
30. Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A.S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., Houlsby, N.: The visual task adaptation benchmark. arXiv preprint **arXiv:1910.04867** (2019) 4
31. Zhang, Y., Zhou, K., Liu, Z.: Neural prompt search. arXiv preprint **arXiv:2206.04673** (2022) 2
32. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 5122–5130. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.544, https://doi.org/10.1109/CVPR.2017.544 1, 4