# Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters

## Supplementary Material

## A. More Implementation Details

We set batch size as 64 for the Multi-domain Task Incremental Learning (MTIL) benchmark and 128 for the Class Incremental Learning (CIL) benchmark. The learning rates are searched among $[10^{-3}, 10^{-4}]$. Label smoothing can substitute the regularization of weight decay and achieve better performance. The label smoothing strength is searched between $\{0.1, 0.2\}$. For CIL, we set weight decay as 0 and label smoothing as 0.0.

## B. Impact of dataset size on expert number

The additional ablation experiments are conducted to explore the optimal number of experts for different task size, and the results are shown in Table 1. The table showcases the impact of dataset size on the optimal number of experts $N_E$ in full-shot setting. We notice that, in general, more tasks require more experts, while simply applying more experts does not always improve accuracy.

## C. Analysis on the Threshold and Different Loss in DDAS

To further analyze the impact of different thresholds ($Thres$) in the Distribution Discriminative Auto-Selector (DDAS), we perform ablation experiments with different thresholds in the methods ("Ours" and "Ours†"), which are shown in Figure 1. The thresholds are searched within the range of $[0.06, 0.07]$. The results show that the performance fluctuation of our method is relatively stable within a certain threshold range. Compared with the method "Ours†", the method "Ours" demonstrates more consistent performance as the threshold changes.

In addition, we conduct ablation experiments on various loss functions for the autoencoder of DDAS, and the results are shown in Table 2. It can be seen that our method achieve the best performance when utilizing the Mean Squared Error (MSE) loss.

## D. More Comparison Results on MTIL

The complete result of the MTIL benchmark with $T$ datasets is a matrix of $T \times T$, where $T$ is the number of incremental tasks. In Table 3 and 4, we present the complete matrices of both "Ours" (trained in 1k iterations) and "Ours†" (trained in 3k iterations) for the MTIL benchmark. In addition, Table 5 and 6 show the results of the full-shot and few-shot MTIL benchmarks in Order-II. The Order-II sequence includes: StanfordCars, Food, MNIST, Oxford-Pet, Flowers, SUN397, Aircraft, Caltech101, DTD, EuroSAT, CIFAR100. As we can see, the proposed method performs favorably against state-of-the-art approaches in terms of three metrics in both settings. Notably, the zero-shot transfer ability of the proposed method closely reaches the upper bound of the pretrained CLIP.

## E. Effectiveness of Router Selection in MoE-Adapters

We visualize the frequency that MoE-Adapters' experts are selected for each incremental task, as shown in Figure 2. As we can see, the activation frequencies of experts are recorded in all visual transformer blocks of CLIP, with 22 experts for each block and $Top\text{-}k$ as 2. The visualization demonstrates the sparsity of the experts activated by our router selection and the cooperation between special experts and shared experts.

| $N_E$ | 4-task | | | 8-task | | | 11-task | | |
|---|---|---|---|---|---|---|---|---|---|
| | Trans. | Avg. | Last | Trans. | Avg. | Last | Trans. | Avg. | Last |
| 2 | 65.8 | 60.7 | 59.0 | 65.9 | 63.2 | 63.2 | 67.3 | 64.1 | 61.5 |
| 4 | 64.9 | 67.5 | 77.1 | 65.0 | 71.2 | 77.9 | 66.5 | 71.1 | 74.1 |
| 8 | 65.1 | 68.3 | 78.3 | 65.4 | 73.7 | 84.9 | 67.4 | 75.7 | 82.4 |
| 16 | 65.3 | 67.7 | 77.9 | 65.5 | 73.9 | 84.9 | 68.0 | 76.4 | 84.6 |
| 20 | 65.5 | 67.4 | 77.0 | 66.6 | 74.6 | 85.8 | 67.6 | 76.0 | 84.2 |

Table 1. Ablation study on the number of experts across different size of dataset.

| Method | full-shot | | | 5-shot | | |
|---|---|---|---|---|---|---|
| | Trans. | Avg. | Last | Trans. | Avg. | Last |
| ZSCL[78] | 68.1 | 75.4 | 83.6 | 65.3 | 66.7 | 67.4 |
| MAE | 68.4 | 73.8 | 77.8 | 68.5 | 70.9 | 70.5 |
| Smooth L1 | 68.3 | 76.9 | 84.9 | 69.0 | 72.9 | 72.6 |
| MSE | 68.9 | 76.7 | 85.0 | 68.9 | 76.3 | 76.1 |

Table 2. Ablation study of different loss in DDAS.

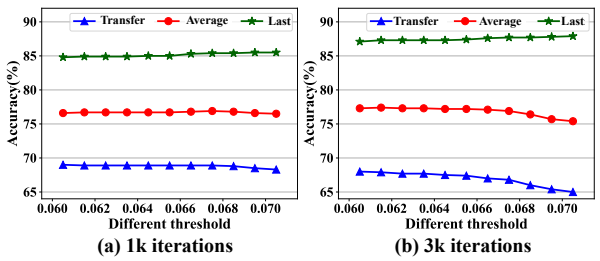


Figure 1. The ablation study of different thresholds in DDAS, and the thresholds are searched within the range of $[0.06, 0.07]$.

| | Aircraft [10] | Caltech101 [5] | CIFAR100 [8] | DTD [2] | EuroSAT [6] | Flowers [11] | Food [1] | MNIST [3] | OxfordPet [12] | Cars [7] | SUN397 [15] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Transfer | | 87.9 | 68.2 | 44.4 | 50.0 | 70.7 | 88.7 | 59.7 | 89.1 | 64.5 | 65.5 | 68.9 |
| Aircraft | 51.5 | 87.9 | 68.2 | 45.1 | 54.6 | 71.3 | 88.8 | 59.5 | 89.1 | 64.5 | 65.3 | |
| Caltech101 | 51.0 | 92.3 | 68.2 | 44.0 | 54.6 | 70.2 | 88.8 | 59.5 | 89.1 | 64.5 | 65.5 | |
| CIFAR100 | 50.0 | 91.5 | 86.7 | 44.2 | 44.4 | 70.8 | 88.8 | 59.8 | 89.1 | 64.5 | 65.5 | |
| DTD | 50.4 | 92.0 | 86.5 | 78.6 | 45.9 | 70.7 | 88.8 | 59.8 | 89.1 | 64.5 | 65.5 | |
| EuroSAT | 50.4 | 91.8 | 86.5 | 78.3 | 96.1 | 70.4 | 88.8 | 59.8 | 89.1 | 64.5 | 65.6 | |
| Flowers | 50.3 | 92.3 | 86.3 | 79.1 | 95.7 | 95.9 | 88.7 | 59.8 | 89.1 | 64.5 | 65.6 | |
| Food | 49.7 | 93.0 | 86.4 | 78.9 | 95.3 | 95.8 | 89.5 | 59.8 | 89.1 | 64.5 | 65.7 | |
| MNIST | 49.7 | 92.7 | 86.3 | 79.0 | 95.5 | 95.6 | 89.5 | 98.3 | 89.1 | 64.5 | 65.6 | |
| OxfordPet | 49.7 | 92.4 | 86.3 | 79.2 | 95.1 | 94.6 | 89.5 | 98.1 | 89.8 | 64.5 | 65.5 | |
| Cars | 49.4 | 92.4 | 86.2 | 78.9 | 94.8 | 94.7 | 89.5 | 98.2 | 89.7 | 81.9 | 65.5 | |
| SUN397 | 49.8 | 92.2 | 86.1 | 78.1 | 95.7 | 94.3 | 89.5 | 98.1 | 89.9 | 81.6 | 80.0 | 85.0 |
| Average | 50.2 | 91.9 | 83.1 | 69.4 | 78.9 | 84.0 | 89.1 | 73.7 | 89.3 | 67.7 | 66.9 | 76.7 |

Table 3. Accuracy (%) of our method (Ours) on the MTIL benchmark with order-I. Each row represents the performance on every dataset of the model trained after the corresponding task. Transfer, Average, and Last metrics are shown in color.

| | Aircraft [10] | Caltech101 [5] | CIFAR100 [8] | DTD [2] | EuroSAT [6] | Flowers [11] | Food [1] | MNIST [3] | OxfordPet [12] | Cars [7] | SUN397 [15] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Transfer | | 87.9 | 68.2 | 42.4 | 41.4 | 68.7 | 88.7 | 59.4 | 89.1 | 64.5 | 64.0 | 67.4 |
| Aircraft | 54.3 | 87.9 | 68.2 | 45.1 | 54.6 | 71.3 | 88.8 | 59.5 | 89.1 | 64.5 | 65.3 | |
| Caltech101 | 54.2 | 92.0 | 68.2 | 40.7 | 54.6 | 67.7 | 88.7 | 59.5 | 89.1 | 64.5 | 63.5 | |
| CIFAR100 | 54.3 | 91.6 | 88.8 | 41.4 | 28.3 | 68.3 | 88.8 | 59.3 | 89.1 | 64.5 | 64.0 | |
| DTD | 54.3 | 91.7 | 88.8 | 80.0 | 28.2 | 68.1 | 88.8 | 59.4 | 89.1 | 64.5 | 64.1 | |
| EuroSAT | 54.3 | 91.7 | 88.8 | 79.9 | 98.0 | 68.1 | 88.8 | 59.4 | 89.1 | 64.5 | 64.0 | |
| Flowers | 54.3 | 91.5 | 88.8 | 79.7 | 98.1 | 97.8 | 88.5 | 59.4 | 89.1 | 64.5 | 63.9 | |
| Food | 54.3 | 91.0 | 88.8 | 80.0 | 98.1 | 97.8 | 89.7 | 59.3 | 89.1 | 64.5 | 63.8 | |
| MNIST | 54.4 | 91.2 | 88.8 | 80.0 | 98.1 | 97.8 | 89.7 | 99.1 | 89.1 | 64.5 | 63.8 | |
| OxfordPet | 54.3 | 90.8 | 88.8 | 79.8 | 98.1 | 97.6 | 89.6 | 99.1 | 89.6 | 64.5 | 63.4 | |
| Cars | 54.2 | 90.8 | 88.8 | 80.2 | 98.1 | 97.5 | 89.6 | 99.1 | 89.5 | 89.2 | 63.8 | |
| SUN397 | 54.3 | 90.8 | 88.8 | 80.3 | 98.1 | 97.5 | 89.6 | 99.1 | 89.5 | 89.2 | 83.8 | 87.4 |
| Average | 54.3 | 91.0 | 85.1 | 69.7 | 77.5 | 84.5 | 89.1 | 73.8 | 89.2 | 69.0 | 65.8 | 77.2 |

Table 4. Accuracy (%) of our method (Ours†) on the MTIL benchmark with order-I. Each row represents the performance on every dataset of the model trained after the corresponding task. Transfer, Average, and Last metrics are shown in color.

## References


[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Proceedings of the European conference on computer vision (ECCV)*, pages 446–461, 2014. 2, 3, 4

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 2, 3, 4

[3] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 2, 3, 4

[4] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model. *arXiv preprint arXiv:2207.09248*, 2022. 3, 4

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incre-

| | Method | Cars [7] | Food [1] | MNIST [3] | OxfordPet [12] | Flowers [11] | SUN397 [15] | Aircraft [10] | Caltech101 [5] | DTD [2] | EuroSAT [6] | CIFAR100 [8] | *Average* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CLIP | Zero-shot | 64.7 | 88.5 | 59.4 | 89.0 | 71.0 | 65.2 | 24.3 | 88.4 | 44.6 | 54.9 | 68.2 | 65.3 |
| | Full Fine-tune | 89.6 | 92.7 | 99.6 | 94.7 | 97.5 | 81.8 | 62.0 | 95.1 | 79.5 | 98.9 | 89.6 | 89.2 |
| | Fine-tune Adapter | 89.1 | 92.9 | 99.2 | 94.1 | 97.0 | 82.7 | 56.8 | 92.6 | 79.0 | 98.4 | 89.4 | 88.3 |
| **Transfer** | Continual-FT | | 85.9 | **59.6** | 57.9 | 40.0 | 46.7 | 11.1 | 70.0 | 30.5 | 26.6 | 37.7 | 46.6 |
| | LwF [9] | | 87.8 | 58.5 | 71.9 | 46.6 | 57.3 | 12.8 | 81.4 | 34.5 | 34.5 | 46.8 | 53.2 |
| | iCaRL [13] | | 86.1 | 51.8 | 67.6 | 50.4 | 57.9 | 11.0 | 72.3 | 31.2 | 32.7 | 48.1 | 50.9 |
| | LwF-VR [4] | | 88.2 | 57.0 | 71.4 | 50.0 | 58.0 | 13.0 | 82.0 | 34.4 | 29.3 | 47.6 | 53.1 |
| | WiSE-FT [14] | | 87.2 | 57.6 | 67.0 | 45.0 | 54.0 | 12.9 | 78.6 | 35.5 | 28.4 | 44.3 | 51.1 |
| | ZSCL [16] | | 88.3 | 57.5 | 84.7 | 68.1 | 64.8 | **21.1** | **88.2** | **45.3** | **55.2** | **68.2** | 64.1 |
| | Ours† | | **88.8** | 59.5 | **89.1** | 69.4 | **65.3** | 15.0 | 87.9 | 43.9 | 54.6 | **68.2** | 64.2**(+0.1)** |
| | Ours | | **88.8** | 59.5 | **89.1** | **69.9** | 64.4 | 18.1 | 86.9 | 43.7 | 54.6 | **68.2** | **64.3**(+0.2) |
| **Average** | Continual-FT | 42.1 | 70.5 | **92.2** | 80.1 | 54.5 | 59.1 | 19.8 | 78.3 | 41.0 | 38.1 | 42.3 | 56.2 |
| | LwF [9] | 49.0 | 77.0 | 92.1 | 85.9 | 66.5 | 67.2 | 20.9 | 84.7 | 44.6 | 45.5 | 50.5 | 62.2 |
| | iCaRL [13] | 52.0 | 75.9 | 77.4 | 74.6 | 58.4 | 59.3 | 11.7 | 79.6 | 42.1 | 43.2 | 51.7 | 56.9 |
| | LwF-VR [4] | 44.9 | 75.8 | 91.8 | 85.3 | 63.5 | 67.6 | 16.9 | 84.9 | 44.0 | 40.6 | 51.3 | 60.6 |
| | WiSE-FT [14] | 52.6 | 79.3 | 91.9 | 83.9 | 63.4 | 65.2 | 23.3 | 83.7 | 45.4 | 40.0 | 48.2 | 61.5 |
| | ZSCL [16] | 81.7 | **91.3** | 91.1 | 91.0 | 82.9 | 72.5 | **33.6** | 89.7 | **53.3** | **62.8** | **69.9** | 74.5 |
| | Ours† | **86.8** | 89.3 | **92.2** | 89.1 | 86.0 | **73.0** | 30.8 | **90.0** | 53.1 | 62.6 | **69.9** | **74.8**(+0.3) |
| | Ours | 84.9 | 89.9 | 89.3 | **91.4** | **86.2** | 72.2 | 33.4 | 89.4 | **53.3** | 61.4 | **69.9** | 74.7(+0.2) |
| **Last** | Continual-FT | 24.0 | 67.3 | 99.1 | 87.4 | 44.3 | 67.0 | 29.5 | 92.3 | 61.3 | 81.0 | **88.1** | 67.4 |
| | LwF [9] | 34.6 | 69.6 | 99.3 | 88.7 | 61.1 | 72.5 | 32.5 | 88.1 | 65.6 | 90.9 | 87.9 | 71.9 |
| | iCaRL [13] | 46.0 | 81.5 | 91.3 | 82.8 | 66.5 | 72.2 | 16.3 | 91.6 | 68.1 | 83.2 | 87.8 | 71.6 |
| | LwF-VR [4] | 27.4 | 61.2 | 99.4 | 86.3 | 60.6 | 70.7 | 23.4 | 88.0 | 61.3 | 84.3 | **88.1** | 68.2 |
| | WiSE-FT [14] | 35.6 | 76.9 | **99.5** | 89.1 | 62.1 | 71.8 | 27.8 | 90.8 | 67.0 | 85.6 | 87.6 | 72.2 |
| | ZSCL [16] | 78.2 | **91.1** | 97.6 | **92.5** | 87.4 | **78.2** | 45.0 | 92.3 | 72.7 | 96.2 | 86.3 | 83.4 |
| | Ours† | 83.7 | 89.0 | 99.2 | 88.7 | 92.9 | 75.9 | 42.6 | 93.1 | 76.6 | **98.6** | 86.4 | **84.2**(+1.8) |
| | Ours | **84.1** | 88.5 | 94.0 | 91.8 | **94.1** | 77.8 | **50.4** | **93.3** | **77.1** | 87.7 | 86.6 | 84.1(+1.7) |

Table 5. Comparison with state-of-the-art methods on MTIL benchmark (Order II) in terms of "Transfer", "Average", and "Last" scores (%). "Ours†" and "Ours" indicate our method trained on 3k and 1k iterations, respectively. We label the best and second methods with **bold** and underline styles. The top block indicates the upper-bound solutions to adapt the CLIP on each task.


mental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 2, 3, 4

[6] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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*, 12(7):2217–2226, 2019. 2, 3, 4

[7] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2, 3, 4

[8] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 3, 4

[9] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 3, 4

[10] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2, 3, 4

[11] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 2, 3, 4

[12] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 2, 3, 4

[13] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 3

[14] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 3, 4

[15] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 2, 3, 4

| | Method | Cars [10] | Food [5] | MNIST [8] | OxfordPet [2] | Flowers [6] | SUN397 [11] | Aircraft [1] | Caltech101 [3] | DTD [12] | EuroSAT [7] | CIFAR100 [15] | *Average* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CLIP | Zero-shot | 64.7 | 88.5 | 59.4 | 89.0 | 71.0 | 65.2 | 24.3 | 88.4 | 44.6 | 54.9 | 68.2 | 65.3 |
| | 5-shot Full Fine-tune | 65.4 | 83.3 | 96.6 | 84.9 | 92.9 | 71.3 | 30.6 | 93.5 | 65.1 | 91.7 | 76.8 | 77.5 |
| | 5-shot Fine-tune Adapter | 68.2 | 87.8 | 90.4 | 89.0 | 94.2 | 72.5 | 29.7 | 90.0 | 63.9 | 81.1 | 75.3 | 76.6 |
| **Transfer** | Continual-FT | | 76.0 | 64.6 | 67.1 | 49.7 | 53.7 | 8.3 | 77.9 | 33.9 | 23.9 | 37.1 | 49.2 |
| | LwF [9] | | 64.2 | 59.1 | 68.1 | 38.4 | 54.9 | 6.7 | 78.0 | 35.5 | 33.5 | 47.4 | 48.6 |
| | LwF-VR [4] | | 80.1 | 55.4 | 77.7 | 50.4 | 61.4 | 9.1 | 83.5 | 40.1 | 31.5 | 54.8 | 54.4 |
| | WiSE-FT [14] | | 77.3 | 60.0 | 76.9 | 54.2 | 58.0 | 11.1 | 81.8 | 37.6 | 31.7 | 48.1 | 53.7 |
| | ZSCL [16] | | 87.3 | **64.8** | 85.3 | 67.9 | 64.5 | **18.9** | 86.6 | 43.6 | 43.2 | 65.7 | 62.8 |
| | Ours | | **88.8** | 59.5 | **89.1** | **71.2** | **65.3** | 18.2 | **87.9** | **44.2** | **54.6** | **68.2** | **64.7**(+1.9) |
| **Average** | Continual-FT | 50.1 | 56.9 | 73.5 | 64.5 | 45.9 | 51.2 | 8.2 | 81.8 | 37.9 | 29.9 | 38.6 | 49.0 |
| | LwF [9] | 64.1 | 55.0 | 79.5 | 69.2 | 55.7 | 58.3 | 10.8 | 81.7 | 41.3 | 39.2 | 47.4 | 54.7 |
| | LwF-VR [4] | 63.3 | 76.9 | 71.4 | 79.1 | 68.9 | 65.0 | 13.4 | 86.0 | 45.7 | 36.3 | 55.3 | 60.1 |
| | WiSE-FT [14] | 59.3 | 64.7 | 77.4 | 70.3 | 51.3 | 58.6 | 10.8 | 84.2 | 42.0 | 38.6 | 49.1 | 55.1 |
| | ZSCL [16] | **70.0** | 85.0 | 79.8 | 86.1 | **79.4** | 68.3 | 21.8 | 88.8 | 48.8 | 49.3 | 66.5 | 67.6 |
| | Ours | 61.2 | **87.0** | **87.3** | **89.1** | 79.3 | **68.5** | **23.4** | **89.4** | **49.9** | **60.8** | **68.8** | **69.5**(+1.9) |
| **Last** | Continual-FT | 35.2 | 28.6 | 58.3 | 51.2 | 14.0 | 46.1 | 5.3 | 89.5 | 47.0 | 52.9 | 53.6 | 42.8 |
| | LwF [9] | 57.1 | 40.1 | 84.1 | 58.1 | 50.5 | 57.6 | 14.3 | 87.9 | 54.7 | 64.0 | 47.0 | 56.8 |
| | LwF-VR [4] | 57.3 | 70.1 | 72.1 | 74.6 | 71.9 | 65.8 | 17.4 | 89.5 | 60.0 | 56.0 | 60.2 | 63.5 |
| | WiSE-FT [14] | 48.1 | 47.7 | 66.9 | 59.8 | 25.0 | 56.1 | 7.4 | 88.5 | 52.2 | 66.8 | 59.4 | 51.8 |
| | ZSCL [16] | **67.4** | 82.7 | 78.7 | 85.7 | 81.3 | 71.2 | 25.0 | **92.5** | 62.0 | 72.2 | 74.4 | 71.8 |
| | Ours | 59.4 | **87.0** | **91.8** | **89.0** | **84.1** | **71.9** | **29.4** | 91.4 | **64.2** | **88.8** | **75.0** | **75.7**(+3.9) |

Table 6. Comparison with state-of-the-art methods on few-shot MTIL benchmark (Order II) in terms of "Transfer", "Average", and "Last" scores (%). Ours converges in 500 iterations on few-shot. We label the best and second methods with **bold** and underline styles. The top block indicates the upper-bound solutions to adapt the CLIP on each task.

[16] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. *arXiv preprint arXiv:2303.06628*, 2023. 3, 4

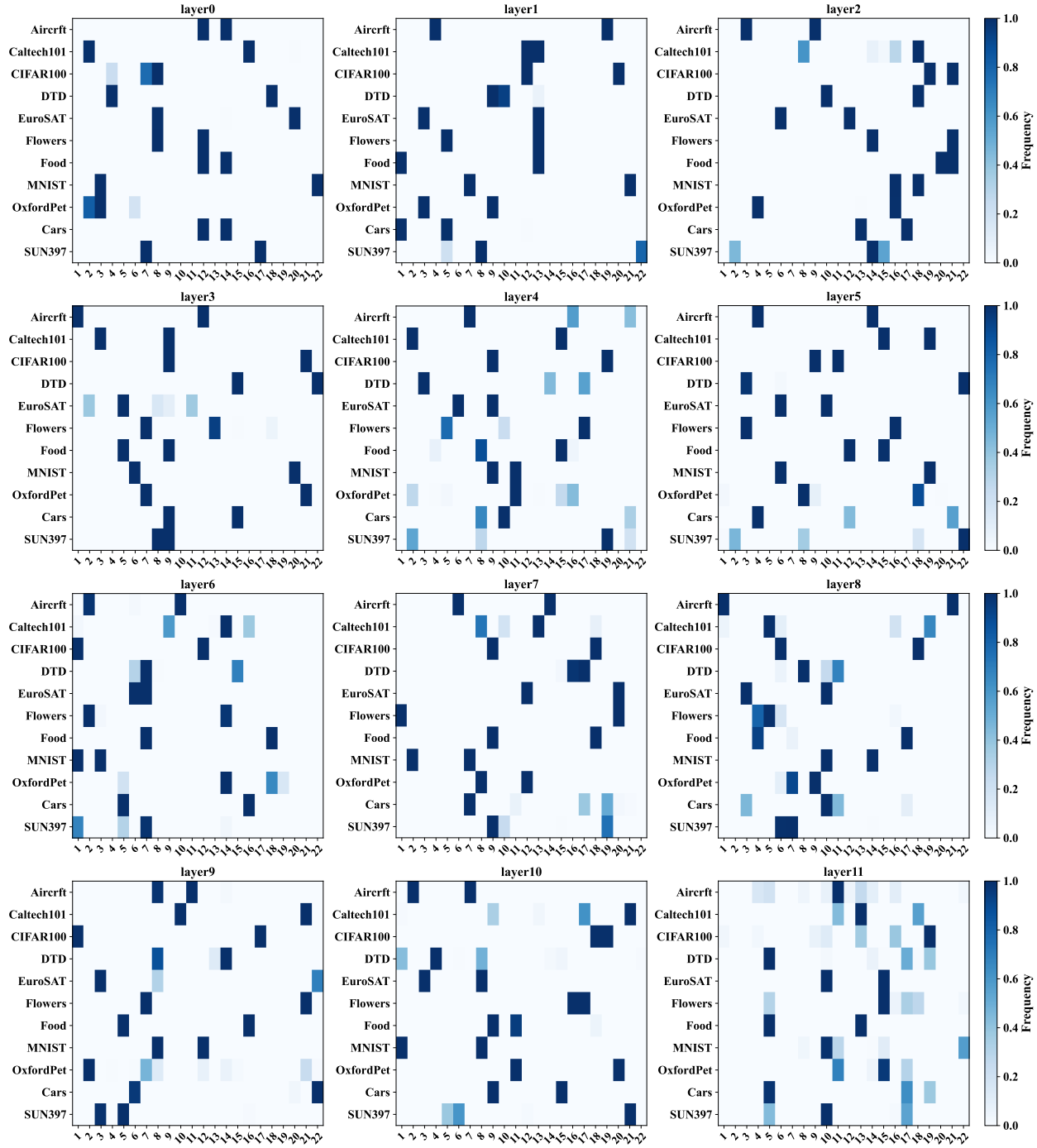


Figure 2. Visualization of the frequency that experts are selected for each task in task incremental learning. The activation frequencies of MoE-Adapters' experts are recorded in all transformer blocks of the visual encoder, with 22 experts and $Top\text{-}K$ as 2. The $y$-axis represents incremental tasks and the $x$-axis represents the experts.