Bridging Domains through Subspace-Aware Model Merging

Supplementary Material

A1. Dataset details

Here, we describe the datasets we use in this worn and include the number of domains and classes for each dataset.

- **PACS [21]** is a small-scale domain-generalization benchmark consisting of four distinct visual styles (domains): Photo, Art painting, Cartoon and Sketch. It contains 7 object classes (dog, elephant, giraffe, guitar, horse, house, person) and about 9,991 images.
- **DomainNet [32]** is a large-scale multi-domain dataset designed for multi-source generalization. It contains 345 object categories spanning six very different domains (clipart, infograph, painting, quickdraw, real (photo), and sketch), with roughly 586k images.
- **ImageNetR [11]** is a rendition-style robustness built from ImageNet labels: renditions such as paintings, cartoons, graffiti, sculptures, toys. It covers 200 ImageNet classes and contains on the order of 30k images.
- **NICOpp [51]** designed to challenge models that learn spurious correlations by grouping images into "contexts" (*e.g.*, "dog on a beach") rather than just broad styles.
- **OfficeHome [41]** it consists of images from 4 different domains: Artistic images, Clip Art, Product images and Real-World images. For each domain, the dataset contains images of 65 object categories found typically in everyday office and home objects.
- **TerraIncognita [2]** it features camera trap images of 10 animal species from four different geographical locations, which serve as the domains. The main objective is specie classification across these domains.
- **FedISIC** comprises dermoscopy images collected from four hospitals. We use same splits of Ogier du Terrail et al. [30] containing 23,247 images from the public training of ISIC2019 [4, 12, 40]. The task involves classifying images into eight melanoma classes, which presents a significant label imbalance. The domains are based on the imaging acquisition device. We have 6 distinct domains.
- **RetinaDomains** composed of a collection of four datasets: Aptos-2019 [17], Messidor [6], IDRDI [33], DDR [22]. The classification task is five Diabetic Retinopathy grading, in which each class corresponds to a grading. We use each dataset as a single domain.

A2. Experimental details

We provide additional details regarding the implementation details and the additional per-domain performances across all datasets and model architectures considered in this work.

A2.1. Implementation details

Our method relies on SVD decomposition, which is defined for any two-dimensional matrix $\Delta w \in \mathbb{R}^{m \times n}$. However, in the model architecture we consider one-dimensional weights also exists. In this case, we follow previous works of Gargiulo et al. [8], Marczak et al. [27] and take the average over all parameters.

We calculate the per-matrix off-diagonal σ_{off} according to Equation 4. Since medical datasets exhibit high sensitivity to off-diagonal noise, we normalize the σ_{off} values of non-medical tasks by the number of merged domains.

A2.2. Fine-tuning details

We fully fine-tune the CLIP's image encoder with a batch size of 128, a learning rate of $1e$-5 coupled with a cosine annealing schedule, and AdamW optimizer [24] with weight decay 0.1 while keeping the text encoder unchanged. We fine-tune, for each domain, a CLIP's visual encoder. We adopt 20 epochs for FedISIC, and RetinaDomains (medical datasets), NICOpp for 15 epochs and 10 epochs otherwise.

A2.3. Multi-task learning performance

We compare SCORE to three methods: Task Arithmetic, TIES, and TSV. The latter is the most cited peer-reviewed method from 2025 in our scorer's suite. We follow the Multi-Task Learning (MTL) scheme across 8 tasks from prior work, using ViT-B-32, ViT-B-16, and ViT-L-14 models. Average accuracy (in %) is reported below. Our method **achieves high IID performance**, while enhancing OOD performance.

Table 4. Average accuracy (in %) of the merged model across 8 MTL datasets across 3 ViT models. Higher is better.

	ViT-B-32	ViT-B-16	ViT-L-14
Task Arithm.	69.98	75.38	82.91
TIES	73.73	78.76	84.14
TSV	83.70	87.23	90.53
SCORE (Ours)	**84.06**	**87.60**	**91.17**

A2.4. Other measures of subspace overlap

We compute per-layer principal angles $\theta \in [0, \pi/2]$ between subspaces as $\theta = arccos(S)$, where $USV^T = SVD(A^T B)$, where $A, B \in \mathbb{R}^{m \times d}$ [19] are subspaces and report the average value in Figure 4. Still, DG presents higher overlap (lower values) measures than MTL.

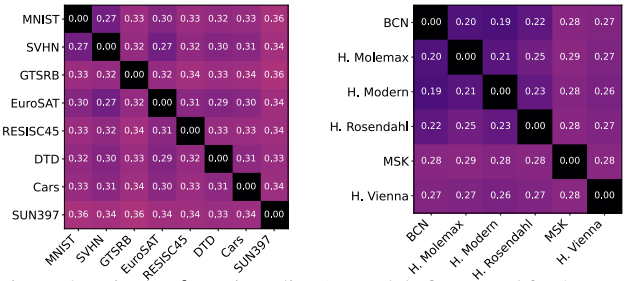

Figure 4. Pairwise θ_{avg} (in radians). Models fine-tuned for 8 MTL datasets(left). Models fine-tuned for 6 domains of FedISIC (right).

A2.5. Per-domain results

We report per-domain performance of the merged model using the leave-one-domain-out evaluation protocol for ViT-B-32 (Figures 5–12), ViT-B-16 (Figures 13–20), and ViT-L-14 (Figures 21–28). Section 5 describes our experimental setup in detail. For completeness, we present results for all model-merging methods we consider in this study, including zero-shot performance, the logits ensemble baseline (see Section 5.2), and the expert performance for each sub-domain.

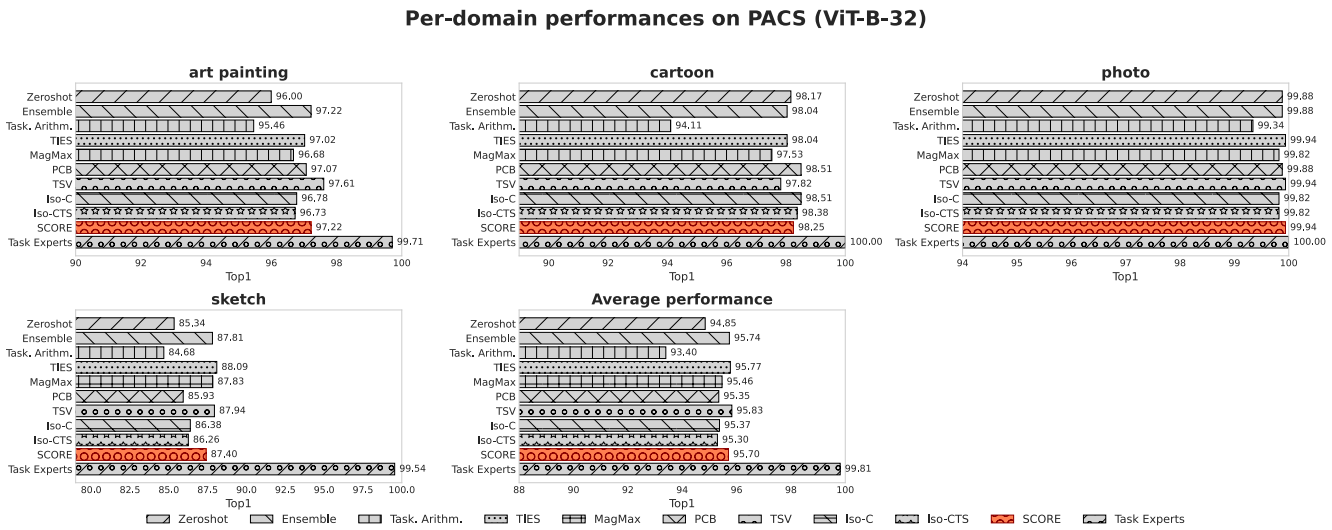

Figure 5. Per-domain results for ViT-B-32 on the PACS dataset for each model merging method in our study.

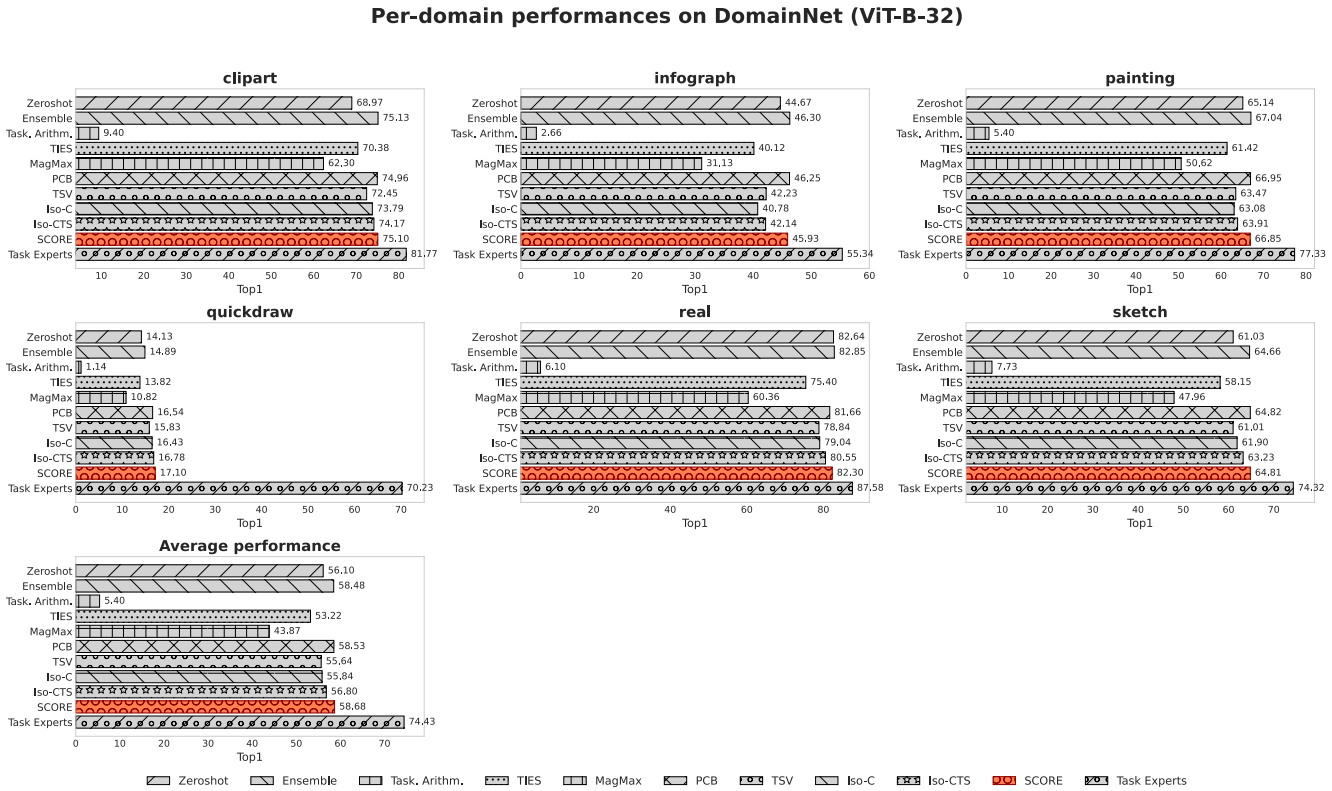

Figure 6. Per-domain results for ViT-B-32 on the DomainNet dataset for each model merging method in our study.

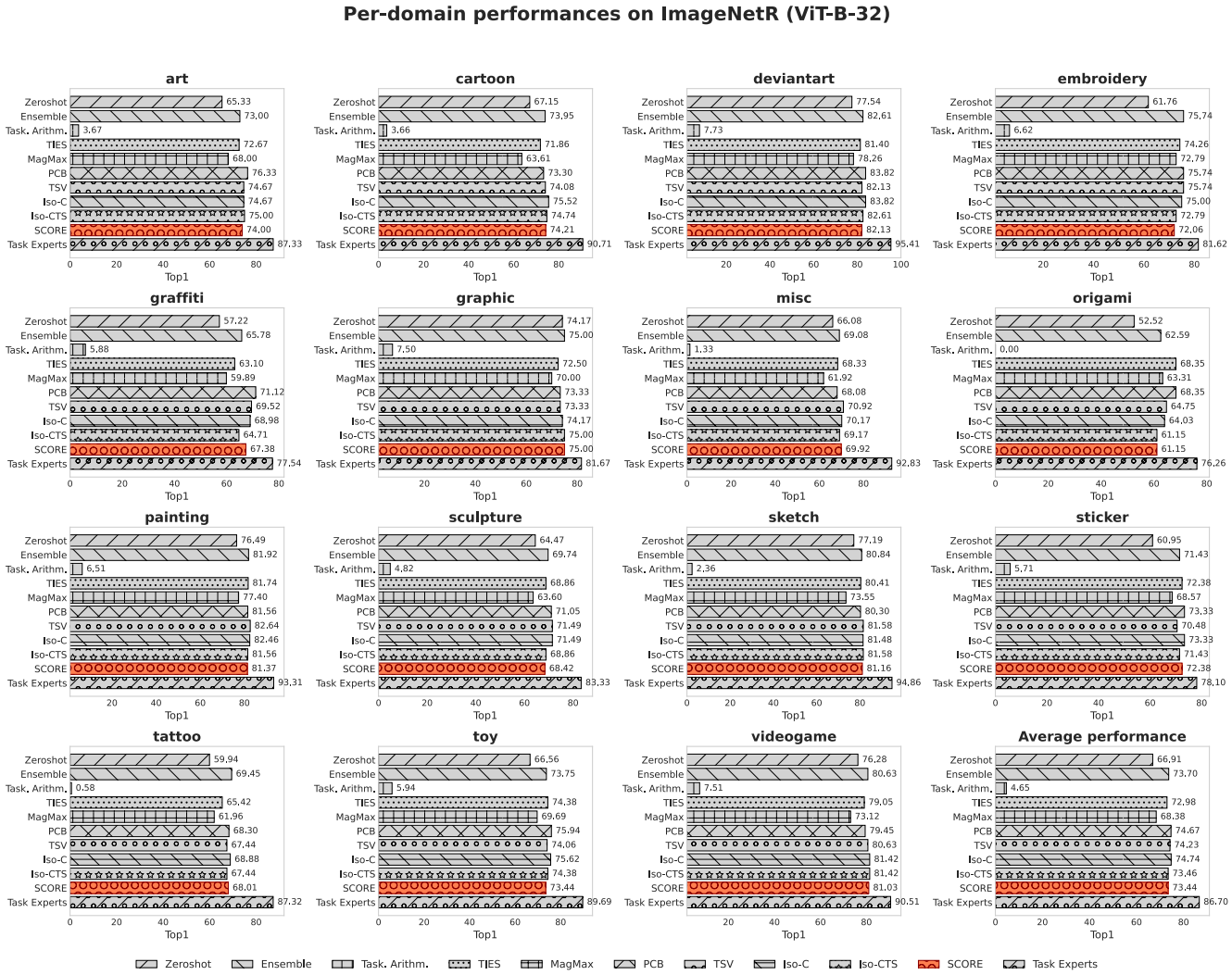

Figure 7. Per-domain results for ViT-B-32 on the ImageNetR dataset for each model merging method in our study.

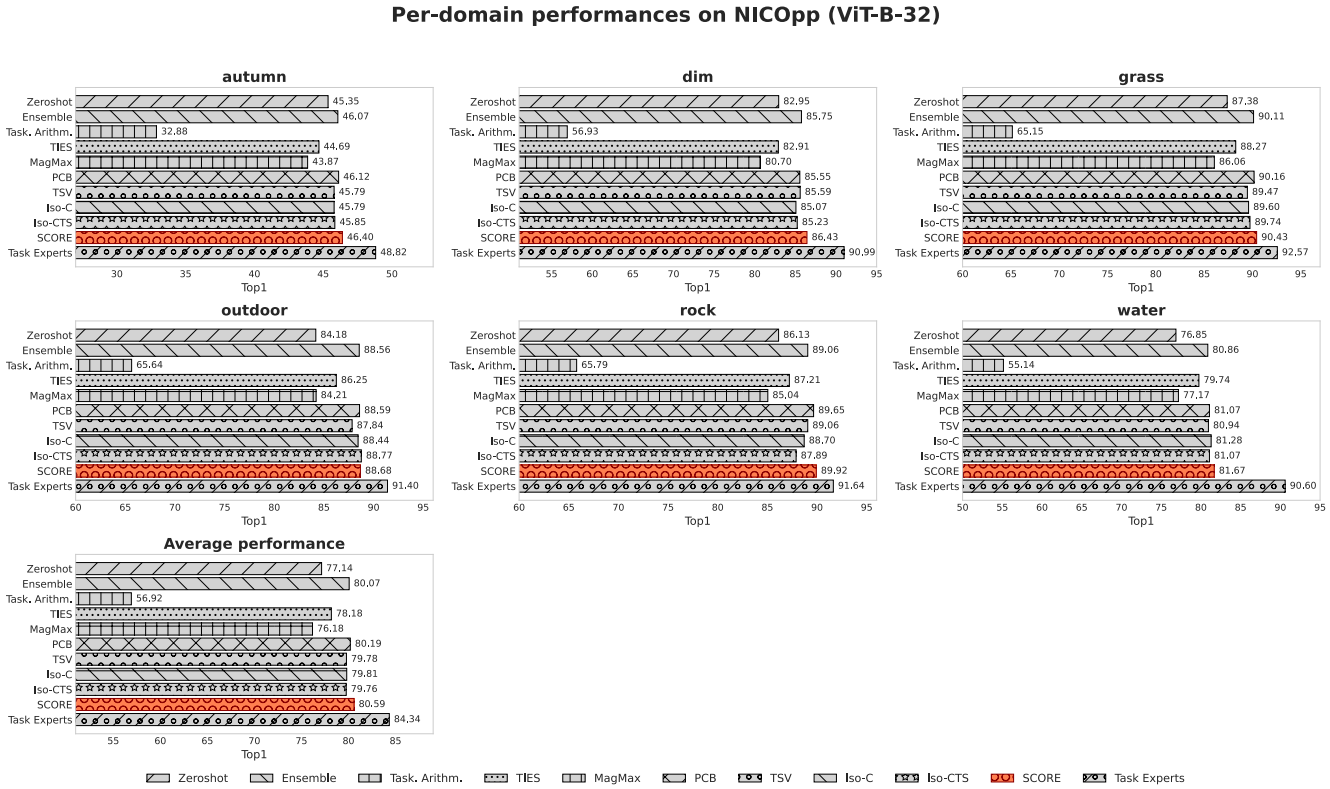

Figure 8. Per-domain results for ViT-B-32 on the NICO++ dataset for each model merging method in our study.

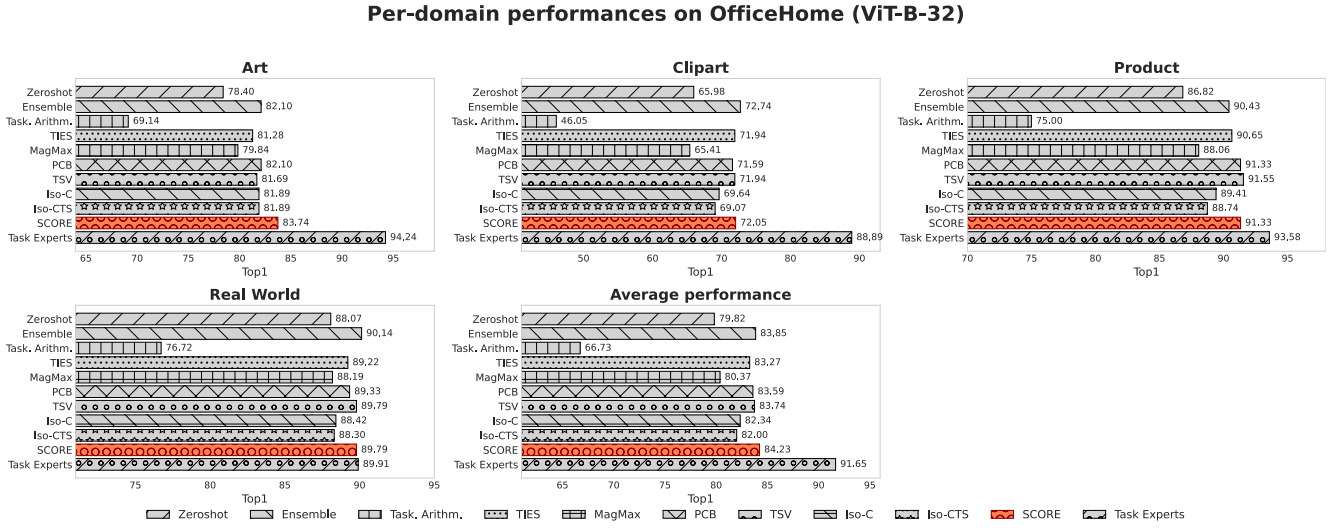

Figure 9. Per-domain results for ViT-B-32 on the OfficeHome dataset for each model merging method in our study.

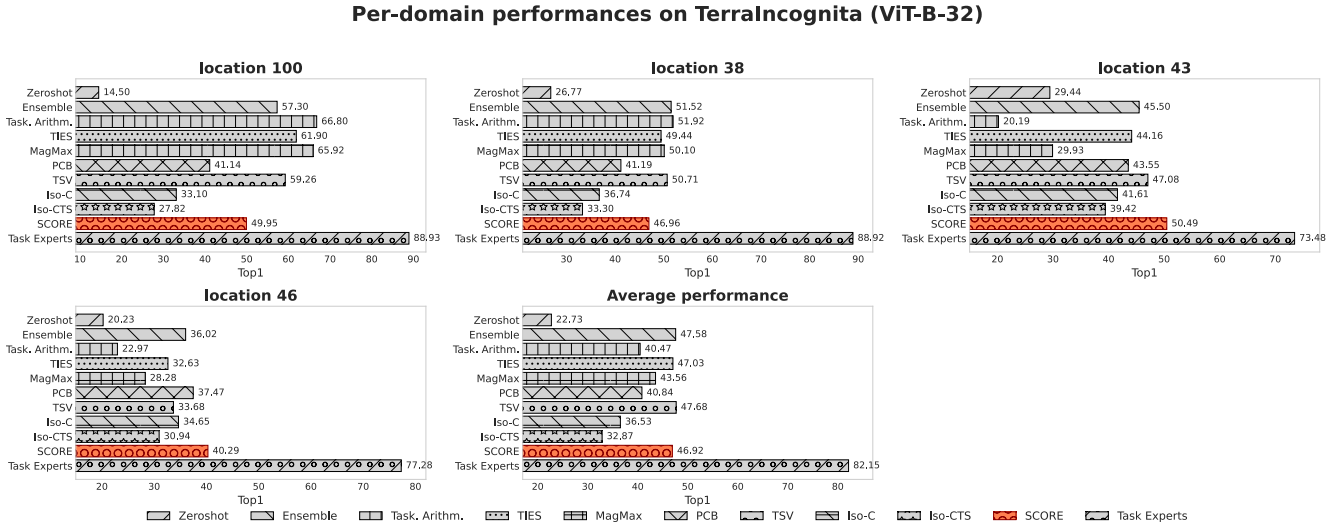

Figure 10. Per-domain results for ViT-B-32 on the TerraIncognita dataset for each model merging method in our study.

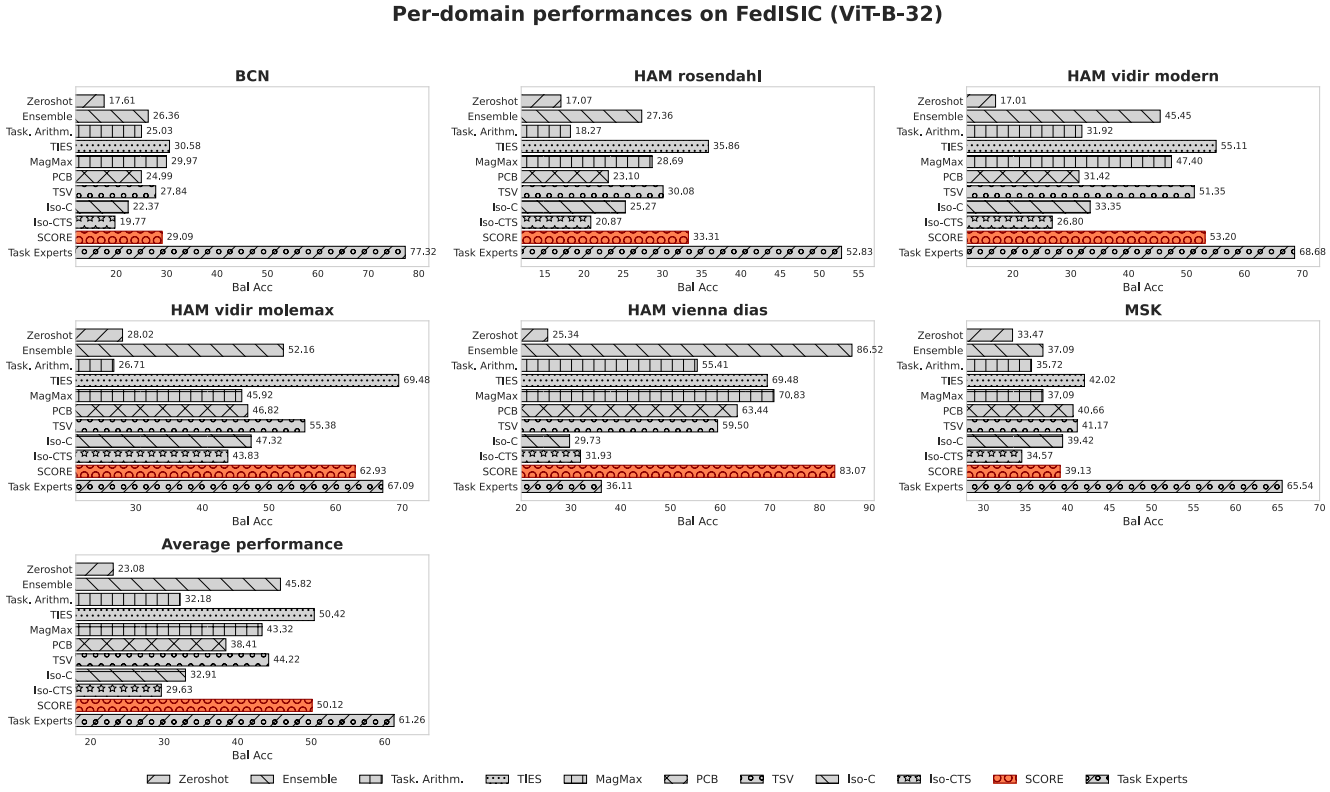

Figure 11. Per-domain results for ViT-B-32 on the FedISIC dataset for each model merging method in our study.

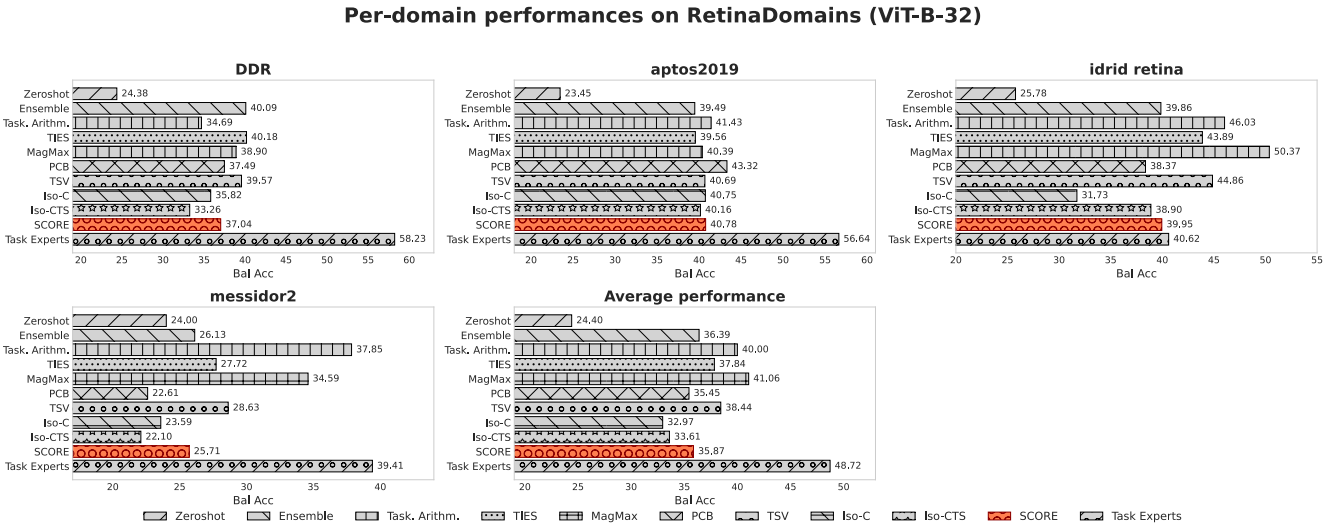

Figure 12. Per-domain results for ViT-B-32 on the RetinaDomains dataset for each model merging method in our study.

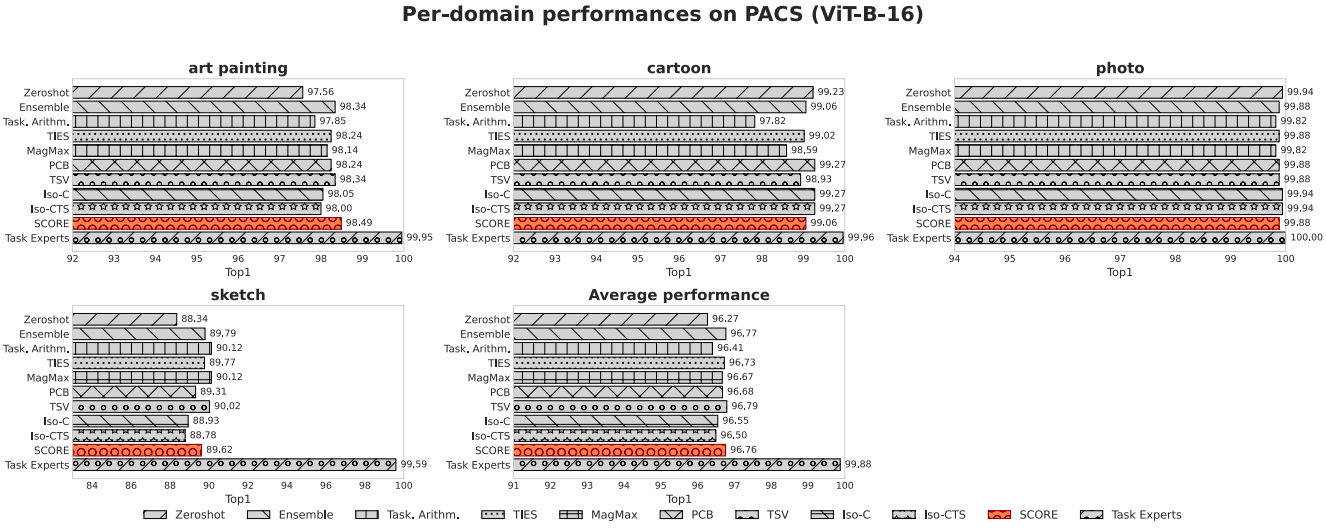

Figure 13. Per-domain results for ViT-B-16 on the PACS dataset for each model merging method in our study.

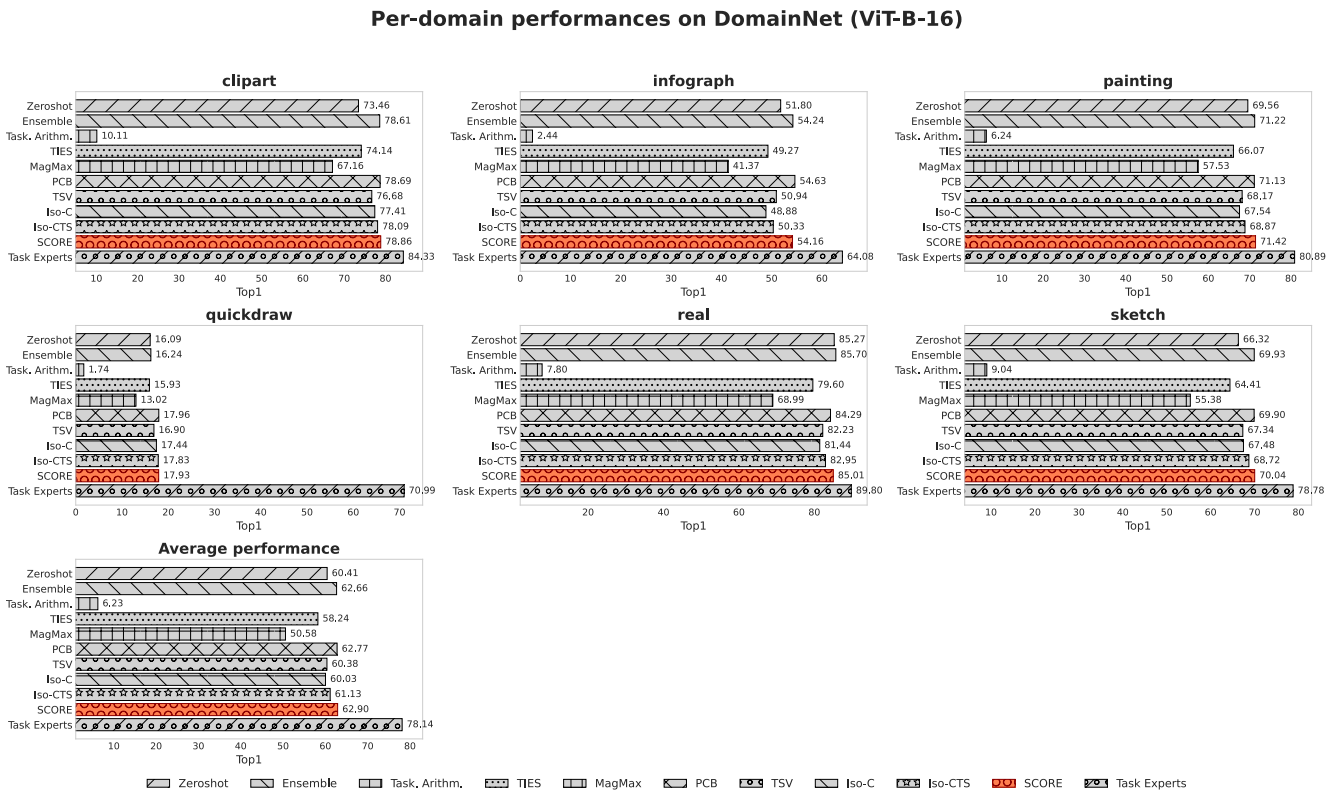

Figure 14. Per-domain results for ViT-B-16 on the DomainNet dataset for each model merging method in our study.

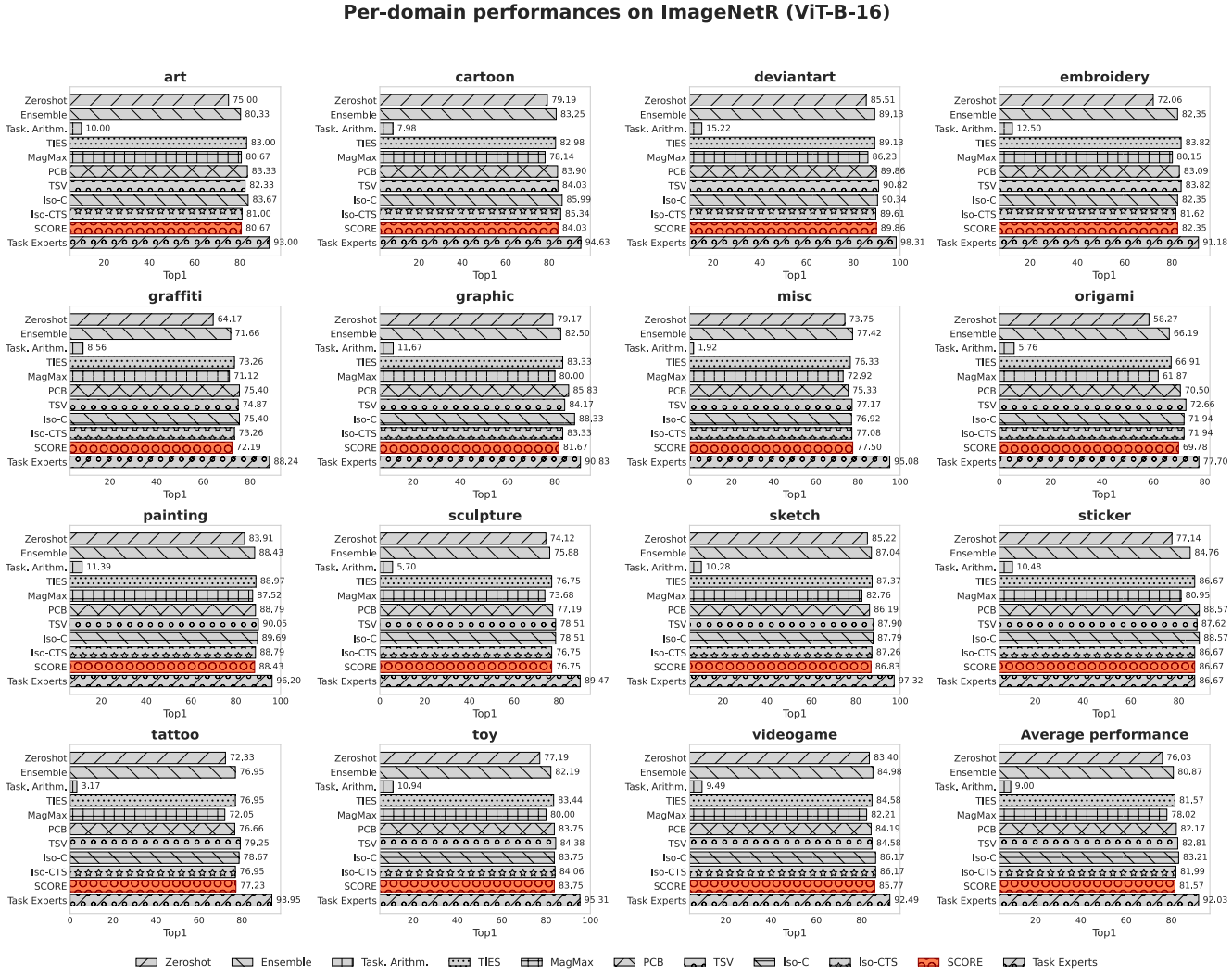

Figure 15. Per-domain results for ViT-B-16 on the ImageNetR dataset for each model merging method in our study.

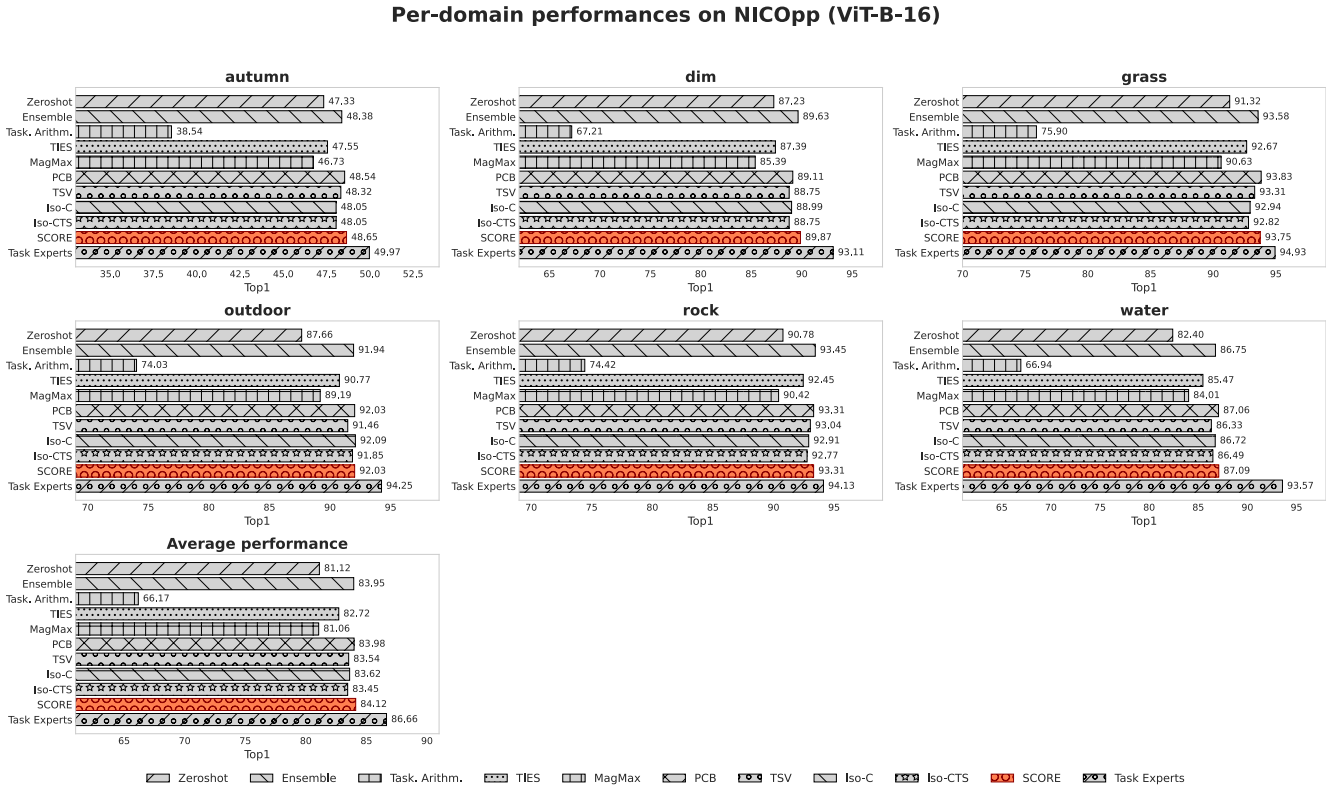

Figure 16. Per-domain results for ViT-B-16 on the NICO++ dataset for each model merging method in our study.

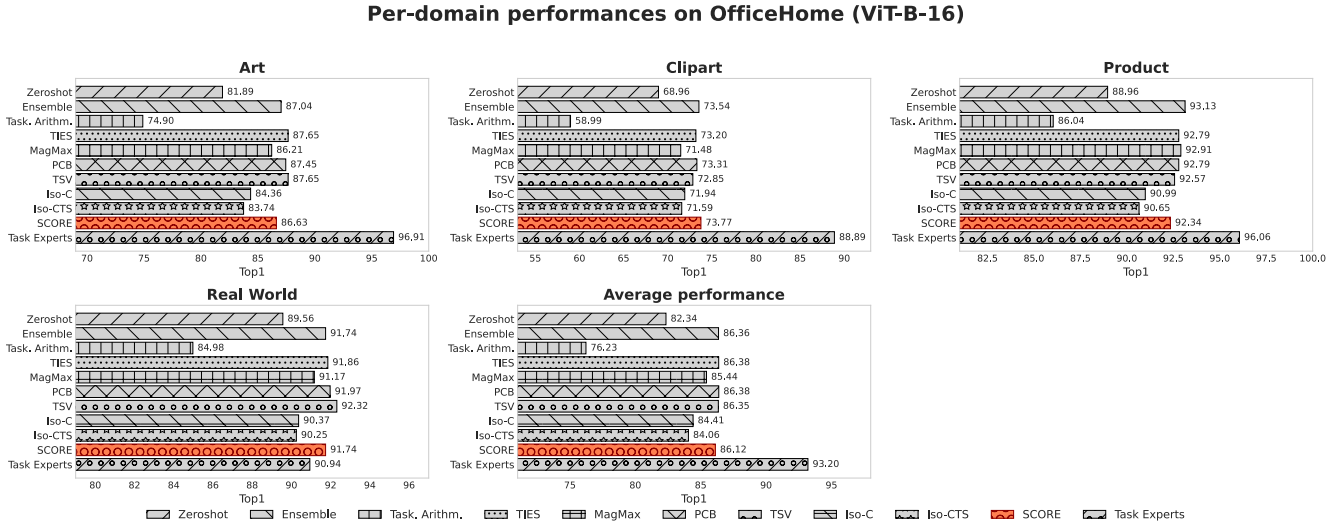

Figure 17. Per-domain results for ViT-B-16 on the OfficeHome dataset for each model merging method in our study.

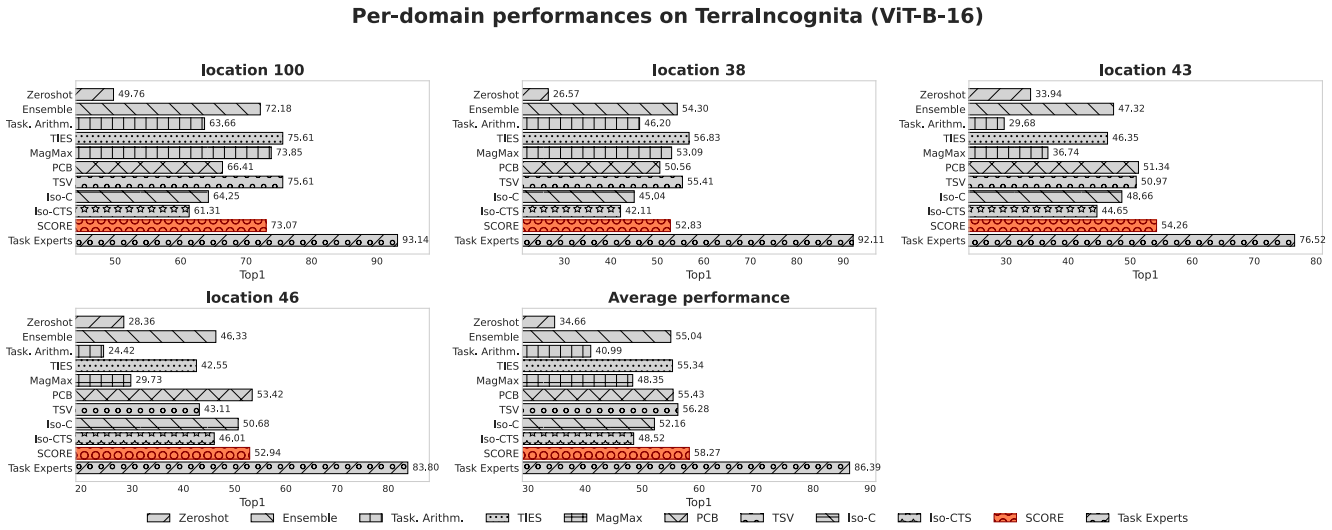

Figure 18. Per-domain results for ViT-B-16 on the TerraIncognita dataset for each model merging method in our study.

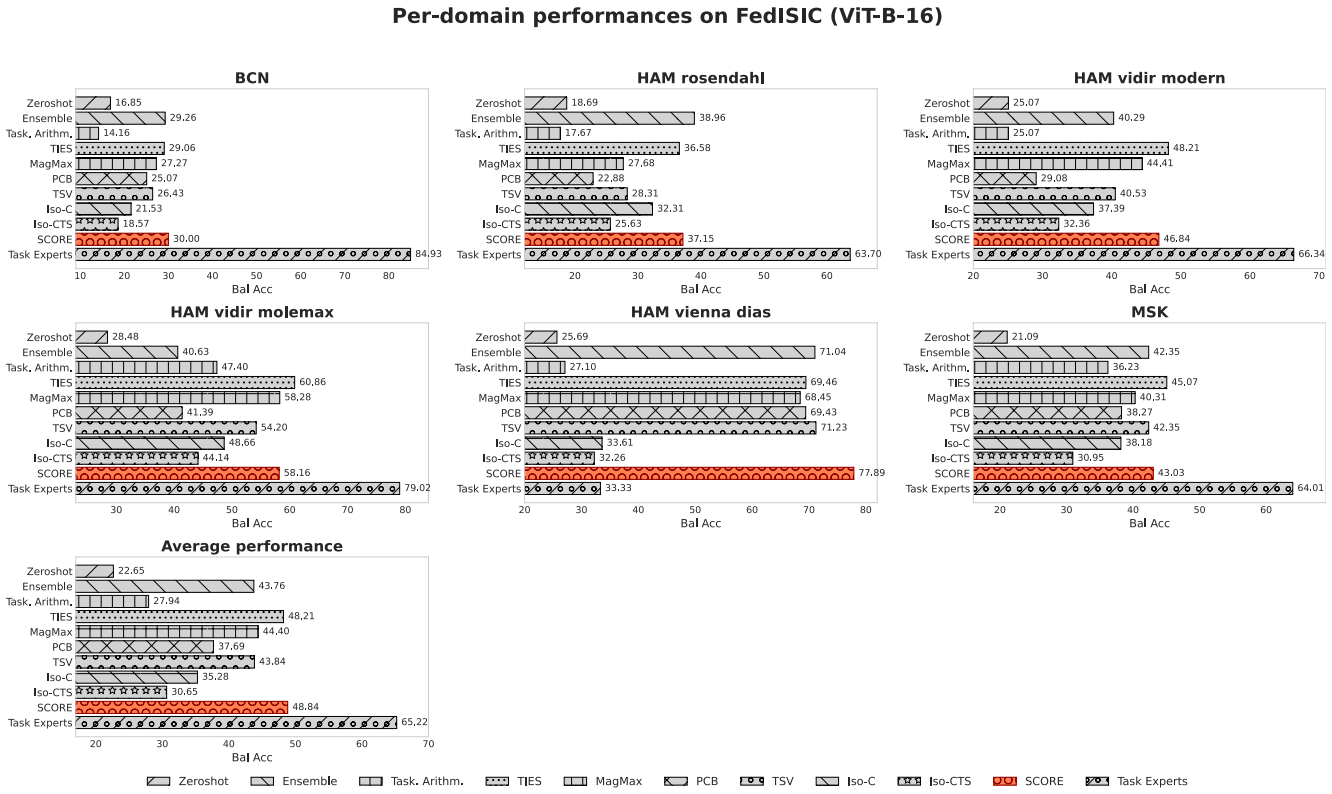

Figure 19. Per-domain results for ViT-B-16 on the FedISIC dataset for each model merging method in our study.

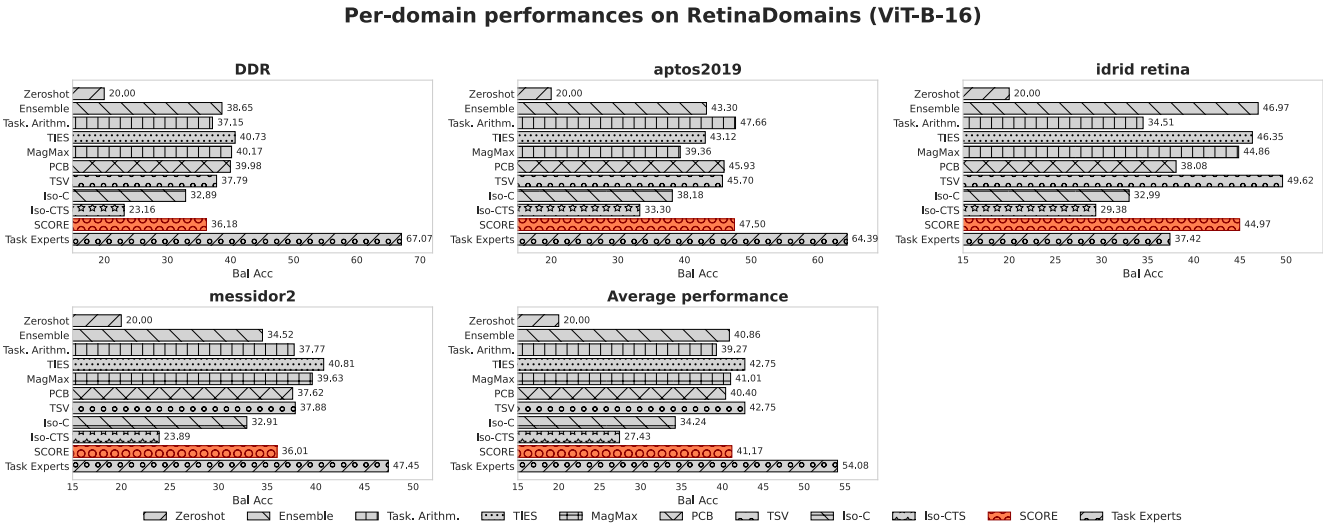

Figure 20. Per-domain results for ViT-B-16 on the RetinaDomains dataset for each model merging method in our study.

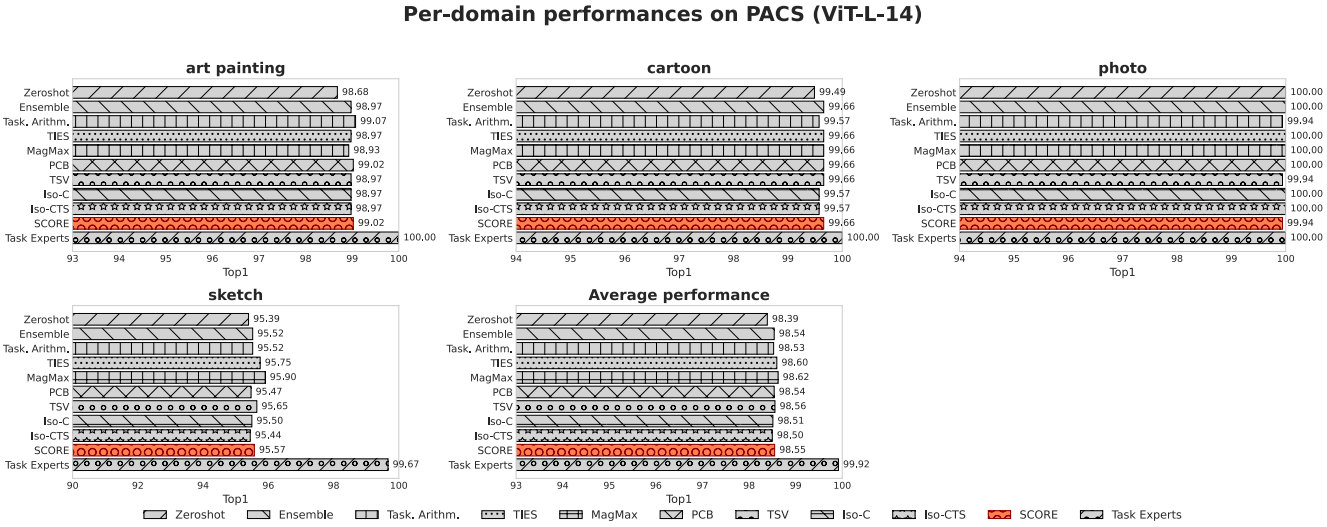

Figure 21. Per-domain results for ViT-L-14 on the PACS dataset for each model merging method in our study.

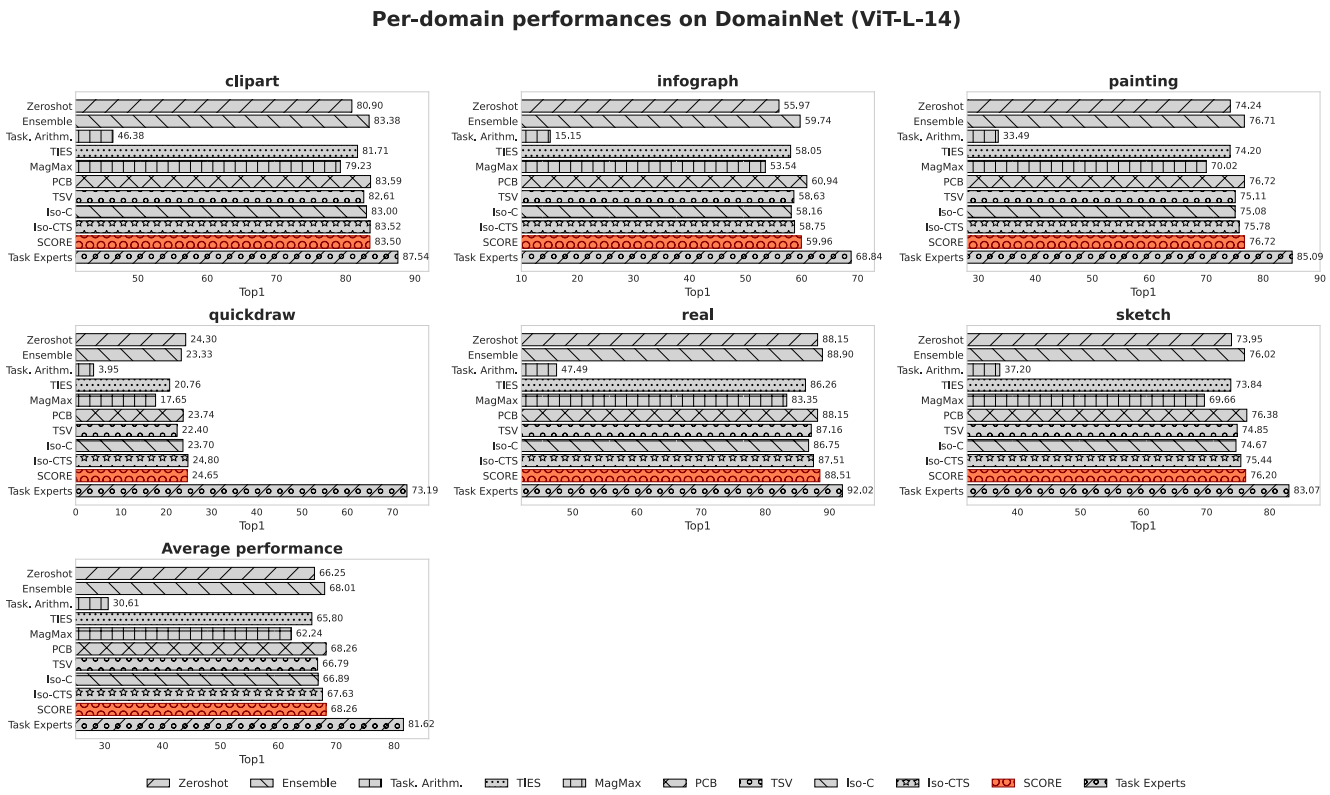

Figure 22. Per-domain results for ViT-L-14 on the DomainNet dataset for each model merging method in our study.

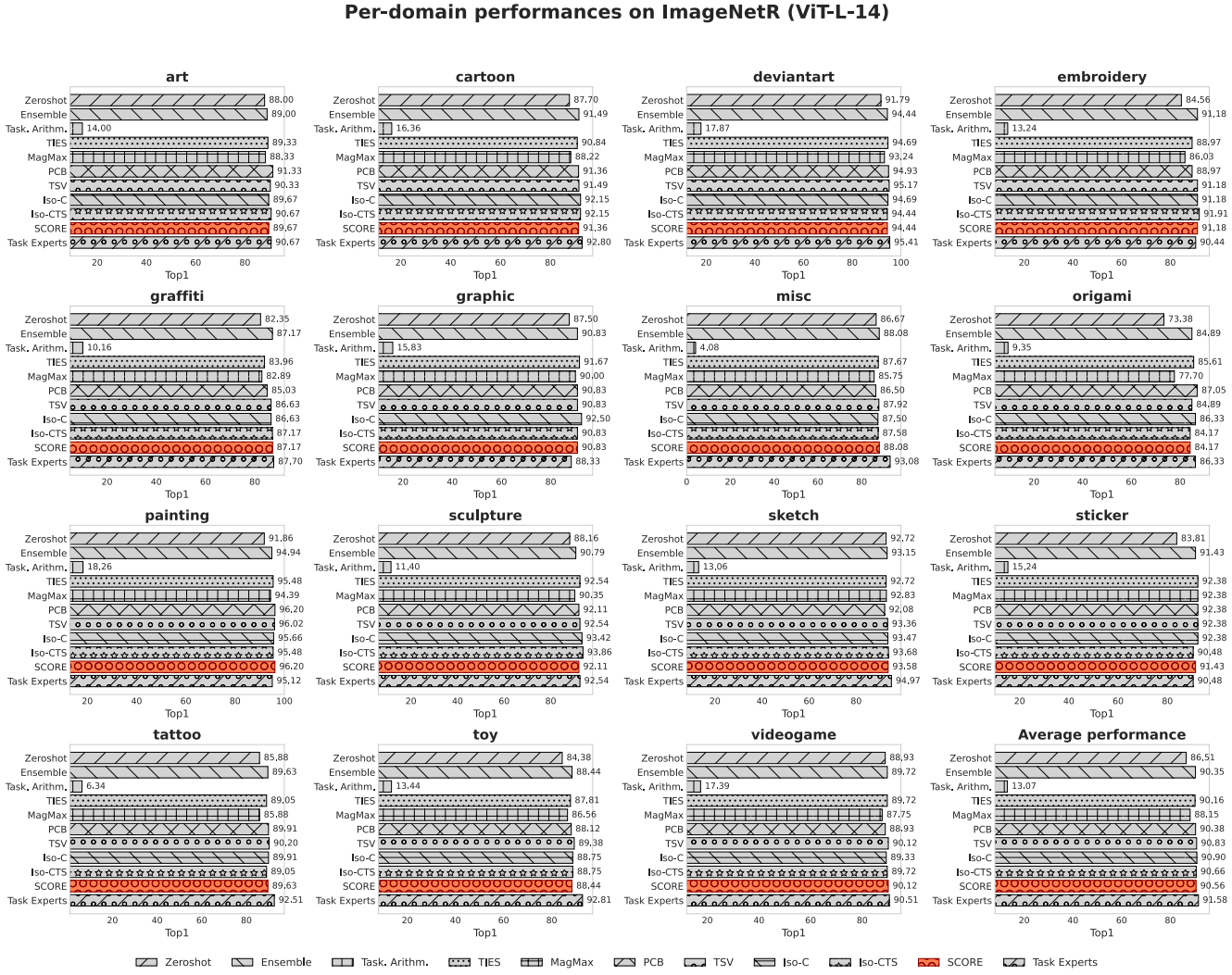

Figure 23. Per-domain results for ViT-L-14 on the ImageNetR dataset for each model merging method in our study.

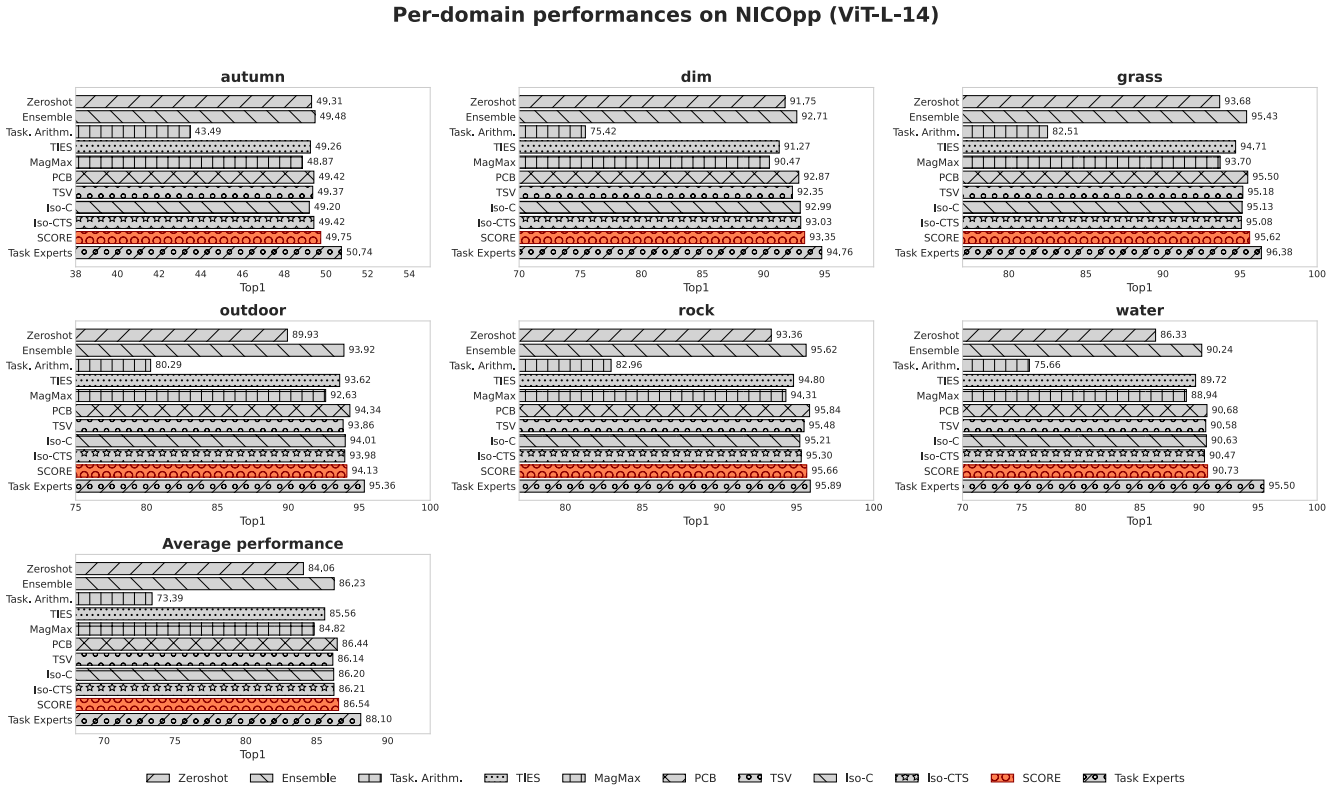

Figure 24. Per-domain results for ViT-L-14 on the NICO++ dataset for each model merging method in our study.

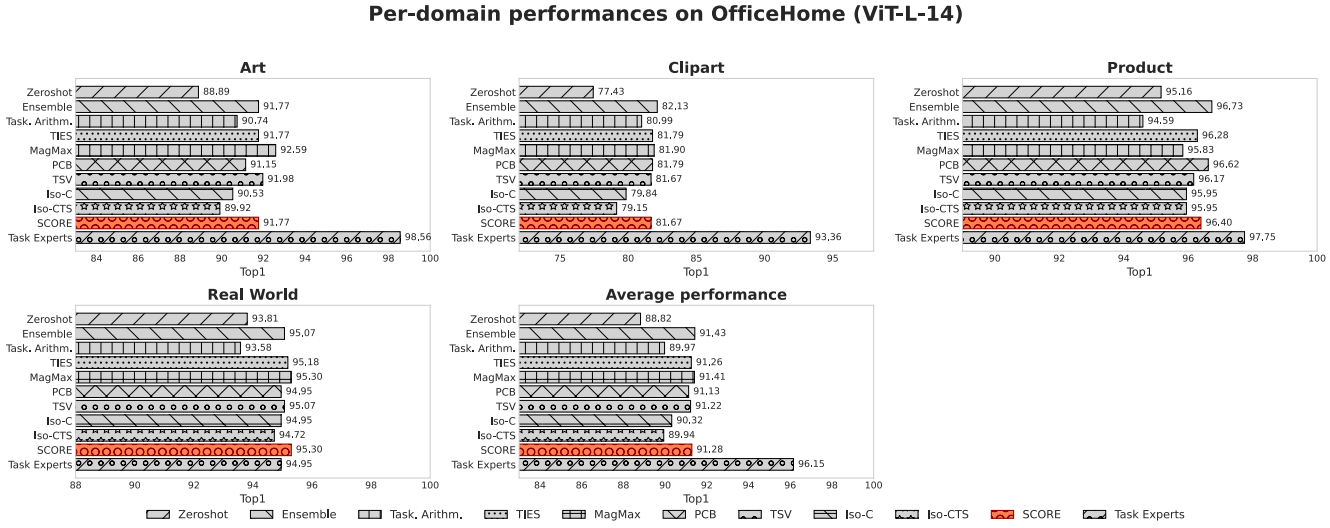

Figure 25. Per-domain results for ViT-L-14 on the OfficeHome dataset for each model merging method in our study.

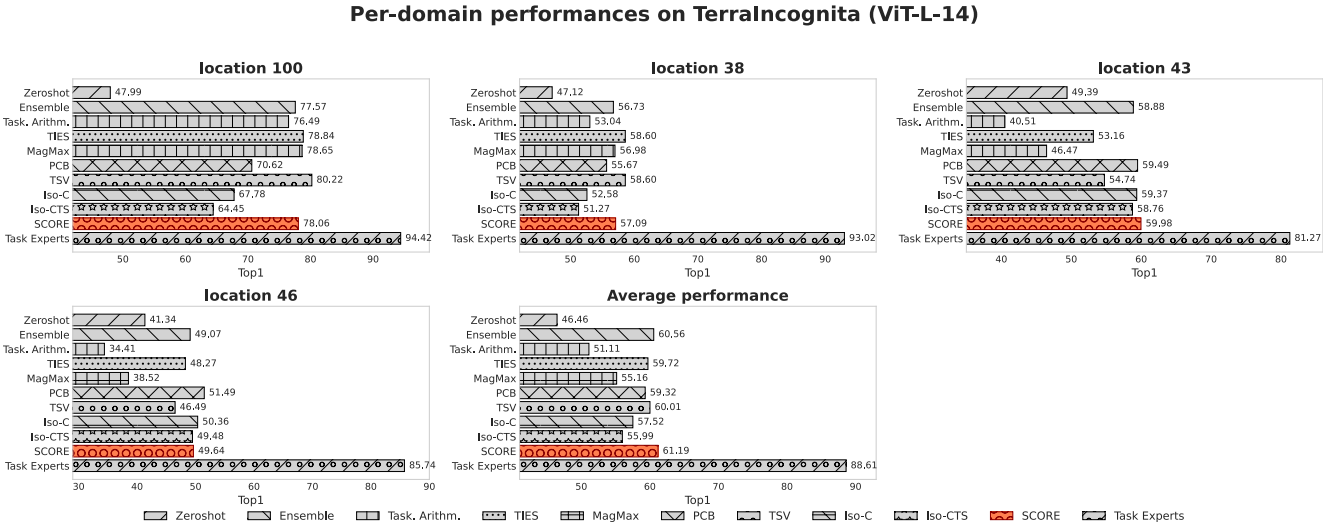

Figure 26. Per-domain results for ViT-L-14 on the TerraIncognita dataset for each model merging method in our study.

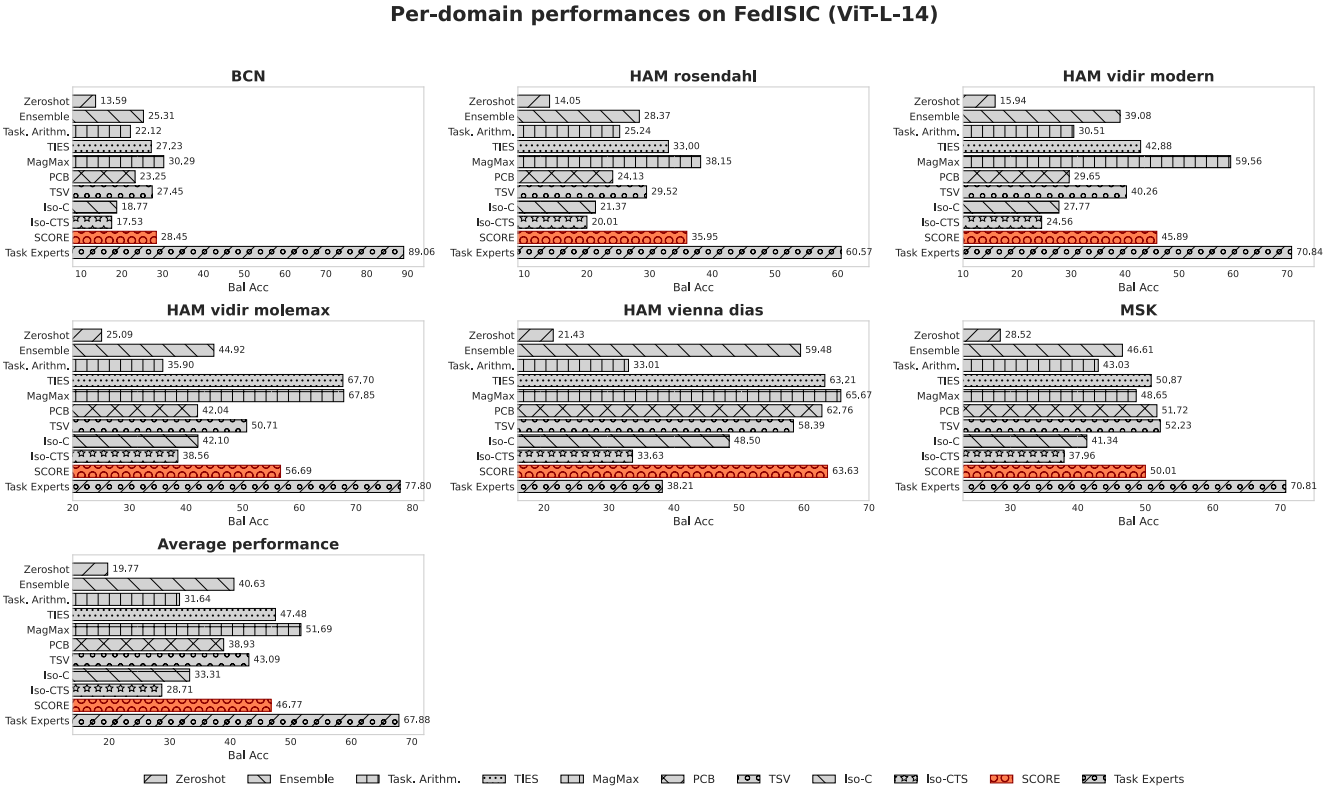

Figure 27. Per-domain results for ViT-L-14 on the FedISIC dataset for each model merging method in our study.

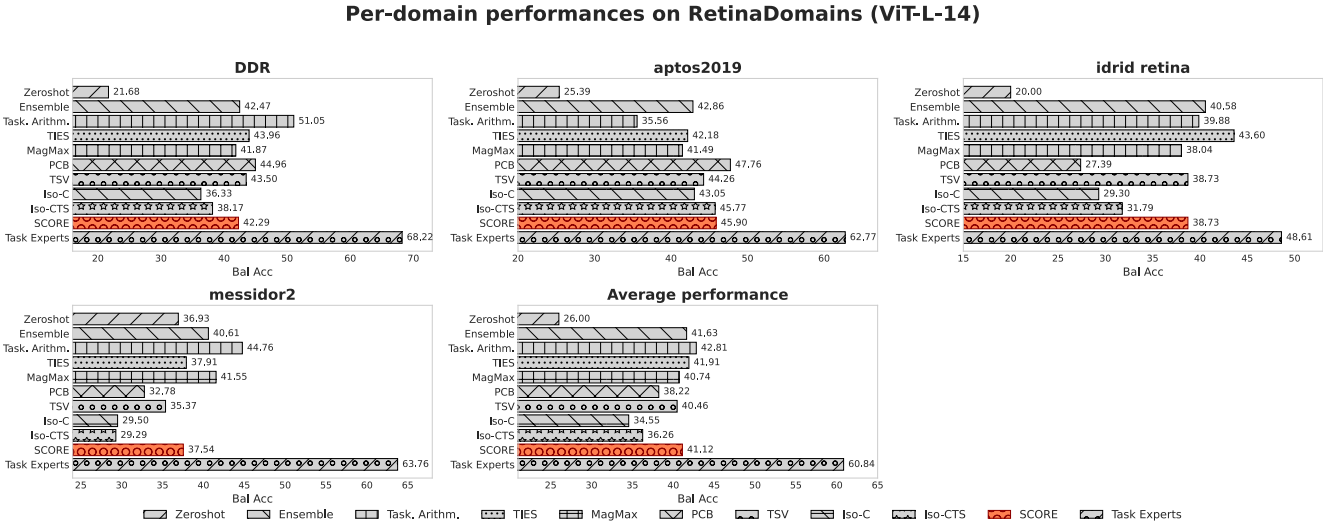

Figure 28. Per-domain results for ViT-L-14 on the RetinaDomains dataset for each model merging method in our study.