

---

# StatsMerging: Statistics-Guided Model Merging via Task-Specific Teacher Distillation

## Supplementary Material

---

Anonymous Author(s)

Affiliation

Address

email

### 1 A Experiment Settings

2 This section presents a comprehensive overview of the datasets, baseline methods, and training  
3 procedures.

4 **Task.** A task is referred to the specific problem or objective that a model is designed to solve. In this  
5 paper, a task is defined as classifying images within a given dataset.

6 **Dataset Details.** This study follows the multi-task model merging protocol from Task Arithmetic  
7 (Ilharco et al., 2023), Ties-Merging (Yadav et al., 2023) and AdaMerging (Yang et al.) on eight image  
8 classification datasets. The details are provided below:

- 9 • **SUN397 (SU)** (Xiao et al., 2016): a scene classification dataset consisting of 397 classes and a  
10 total of 108,754 images, with each class containing a minimum of 100 images.
- 11 • **Stanford Cars (CA)** (Krause et al., 2013): a car classification benchmark dataset comprising  
12 196 categories and 16,185 images in total. For each category, the dataset is evenly divided into  
13 training and test sets in a 1:1 ratio.
- 14 • **RESISC45 (RE)** (Cheng et al., 2017): a remote sensing image scene classification benchmark  
15 with 45 scene classes and 31,500 images. Approximately 700 images are included in each class.
- 16 • **EuroSAT (EU)** (Helber et al., 2019): a 10-class satellite image classification dataset with 27,000  
17 labeled and geo-referenced images.
- 18 • **SVHN (SV)** (Netzer et al., 2011): a real-world digit classification dataset derived from house  
19 numbers in Google Street View images. This datasets consists of 10 classes with 73,257 training  
20 samples and 26,032 test samples. Additional 531,131 samples are available for training.
- 21 • **GTSRB (GT)** (Stallkamp et al., 2011): a traffic sign classification dataset consisting of 43 classes  
22 and more than 50,000 samples in total.
- 23 • **MNIST (MN)** (LeCun et al., 1998): a benchmark dataset for image classification, containing  
24 grayscale images of handwritten digits across 10 classes. It includes 60,000 training and 10,000  
25 test images, with a balanced number across classes.
- 26 • **DTD (DT)** (Cimpoi et al., 2014): a texture classification dataset consisting of 47 classes and a  
27 total of 5,640 images, with approximately 120 images per class.

28 **Baseline Details.** We evaluate performance using eight comparison baselines and four alternative  
29 configurations of our method.

- 30 • **Individual:** Each task is handled by an independently fine-tuned model with no interference  
31 between tasks. However, this approach cannot perform multiple tasks simultaneously.
- 32 • **Traditional MTL:** This approach aggregates the original training data from all tasks to train  
33 a single multi-task model. It serves as a reference *upper bound* for evaluating model merging  
34 performance.

- 35 • **Weight Averaging**: A simple model merging technique that averages the parameters of multiple  
36 models directly. It is typically considered a *lower bound* for model merging performance.
- 37 • **Fisher Merging** (Matena and Raffel, 2022): This method computes the Fisher Information Matrix  
38 to assess parameter importance, guiding the model merging process based on these importance  
39 scores.
- 40 • **RegMean** (Jin et al., 2023): Introduces a regularization constraint during merging, enforcing the  
41  $L_2$  distance between the merged model and individual models to remain small.
- 42 • **Task Arithmetic** (Ilharco et al., 2023): This method is the first to propose the concept of “task  
43 vectors” and merges these vectors into a pre-trained for model merging.
- 44 • **Ties-Merging** (Yadav et al., 2023): Addresses task conflict in Task Arithmetic (Ilharco et al., 2023)  
45 by removing redundant parameters and resolving sign conflicts through a three-step procedure:  
46 Trim, Elect Sign, and Disjoint Merge.
- 47 • **EMR-MERGING** (Huang et al., 2024): This approach is a tuning-free method that merges models  
48 in three steps, by selecting a unified parameter sign (Elect), aligning task-specific parameters via  
49 masking (Mask), and adjusting their magnitudes with task-specific scaling factors (Rescale).
- 50 • **AdaMerging** (Yang et al.): Builds on Task Arithmetic (Ilharco et al., 2023) by employing an  
51 unsupervised method to automatically learn merging coefficients for each task vector.
- 52 • **AdaMerging++** (Yang et al.): An extension of Ties-Merging (Yadav et al., 2023) that uses an  
53 unsupervised approach to learn task-specific merging coefficients.
- 54 • **StatsMerging (Ours)**: A lightweight learning-based method guided by the weight distribution  
55 statistical features (stats) of task-specific pre-trained weight models, including the mean, variance,  
56 magnitude and singular values. This method employs *StatsMergeLearner* learn stats by knowledge  
57 distillation from task-specific teachers without manual labels.
- 58 • **StatsMerging++ (Ours)**: A more extensively trained version of *StatsMerging*.

#### 59 **Training Details.**

- 60 • **Task-Specific Teacher**: For each task, we utilize its corresponding **Individual** model as the  
61 **Teacher**.

62 Code is available at <https://github.com/statsmerging/statsmerging>.

## 63 B Extended Experiments

### 64 B.1 Merging Performance

65 Extended experimental merging results are presented in Table 1. Results for Pre-Trained models,  
66 Individual models, and those trained using Traditional MTL are listed above the double horizontal  
67 lines. Below these lines, the comparison is organized into three groups: Task-wise methods appear  
68 first, followed by Layer-wise approaches, and finally the Parameter-wise method. Notably, while  
69 finer granularity is generally associated with improved merging performance (Yang et al.), our **LW**  
70 **StatsMerging++**, operating at the Layer-wise level, surpasses EMR-Merging (Huang et al., 2024),  
71 which is based on the finer Parameter-wise granularity.

Table 1: Multi-task merging performance (Avg Acc %) when merging ViT-B/32 models on eight tasks. Results of our method *StatsMerging* are shaded in gray. Bold and underscore indicate the highest and second-highest scores within the merging group below the double rules in each column, respectively. GL: Granularity Level. TW: Task-wise. LW: Layer-wise. PW: Parameter-wise.

Method	SU	CA	RE	EU	SV	GT	MN	DT	Avg Acc
Pre-Trained	62.3	59.7	60.7	45.5	31.4	32.6	48.5	43.8	48.0
Individual	75.3	77.7	96.1	99.7	97.5	98.7	99.7	79.4	90.5
Traditional MTL	73.9	74.4	93.9	98.2	95.8	98.9	99.5	77.9	88.9
<b>Task-wise</b>									
Weight Averaging	65.3	63.4	71.4	71.7	64.2	52.8	87.5	50.1	65.8
Task Arithmetic	55.2	54.9	66.7	78.9	80.2	69.7	97.3	50.4	69.1
Fisher Merging	68.6	69.2	70.7	66.4	72.9	51.1	87.9	59.9	68.3
RegMean	65.3	63.5	75.6	78.6	78.1	67.4	93.7	52.0	71.8
Ties-Merging	59.8	58.6	70.7	79.7	86.2	72.1	98.3	54.2	72.4
TW AdaMerging	58.0	53.2	68.8	85.7	81.1	84.4	92.4	44.8	71.1
TW AdaMerging++	60.8	56.9	73.1	83.4	87.3	82.4	95.7	50.1	73.7
<b>TW StatsMerging (Ours)</b>	61.3	70.0	74.2	85.2	87.5	82.5	96.2	54.2	76.4
<b>Layer-wise</b>									
LW AdaMerging	64.5	68.1	79.2	93.8	87.0	91.9	97.5	59.1	80.1
LW AdaMerging++	66.6	68.3	82.2	94.2	89.6	89.0	98.3	60.6	81.1
<b>LW StatsMerging (Ours)</b>	67.4	<u>74.1</u>	82.9	91.1	89.8	94.7	98.3	<u>77.5</u>	84.5
<b>LW StatsMerging++ (Ours)</b>	<u>74.1</u>	<b>74.2</b>	<u>93.4</u>	<u>98.7</u>	<u>96.7</u>	<b>98.5</b>	<u>98.8</u>	<b>77.7</b>	<b>89.0</b>
<b>Parameter-wise</b>									
EMR-MERGING	<b>75.2</b>	72.8	<b>93.5</b>	<b>99.5</b>	<b>96.9</b>	<u>98.1</u>	<b>99.6</b>	74.4	<u>88.7</u>

## 72 B.2 Robustness Evaluation

73 We evaluate the robustness of *StatsMerging* against Task Arithmetic (Ilharco et al., 2023) and  
 74 AdaMerging (Yang et al.) under three image corruption scenarios: Motion Blur, Impulse Noise, and  
 75 Gaussian Noise. The corrupted test sets are constructed following the protocols outlined in (Yang  
 76 et al.; Hendrycks and Dietterich, 2019). We assess performance on four datasets: Stanford Cars  
 77 (CA) (Krause et al., 2013), EuroSAT (EU) (Helber et al., 2019), RESISC45 (RE) (Cheng et al., 2017),  
 78 and GTSRB (GT) (Stallkamp et al., 2011). Results are reported in Table 2. Overall, *StatsMerging*  
 79 consistently outperforms the baselines. On the clean test set, it achieves a 2.4% accuracy improvement  
 80 over AdaMerging. Under corrupted conditions, *StatsMerging* yields performance gains of 3.1%,  
 81 6.3%, and 4.3% for Motion Blur, Impulse Noise, and Gaussian Noise, respectively.

Table 2: Robustness results when merging ViT-B/32 models on four tasks. *StatsMerging*: shaded in gray. Bold: top score. Values are reported in %.

Method	CA	EU	RE	GT	Avg Acc
<b>Clean Test Set</b>					
Task Arithmetic	66.9	94.7	82.6	75.1	79.8
AdaMerging	73.7	96.1	85.8	96.3	88.0
<i>StatsMerging</i>	<b>75.6</b>	<b>96.3</b>	<b>92.1</b>	<b>97.6</b>	<b>90.4 (+2.4)</b>
<b>Motion Blur</b>					
Task Arithmetic	65.3	68.1	80.0	64.2	69.4
AdaMerging	71.2	74.6	82.7	94.1	80.6
<i>StatsMerging</i>	<b>73.5</b>	<b>76.9</b>	<b>89.2</b>	<b>95.2</b>	<b>83.7 (+3.1)</b>
<b>Impulse Noise</b>					
Task Arithmetic	62.1	49.1	72.7	40.4	56.1
AdaMerging	67.2	30.8	75.9	77.5	62.8
<i>StatsMerging</i>	<b>70.4</b>	<b>50.4</b>	<b>77.6</b>	<b>78.1</b>	<b>69.1 (+6.3)</b>
<b>Gaussian Noise</b>					
Task Arithmetic	63.6	55.4	75.9	49.4	61.1
AdaMerging	69.9	41.2	80.6	76.0	66.9
<i>StatsMerging</i>	<b>71.2</b>	<b>53.6</b>	<b>82.1</b>	<b>78.0</b>	<b>71.2 (+4.3)</b>

### 82 B.3 Label Type and Loss Function Analysis

83 In this section, we analyze the performance of training *StatsMergeLearner* on two types of pseudo  
 84 labels: (1) Soft Pseudo Labels, and (2) Hard Pseudo Labels, the former of which is commonly  
 85 employed in knowledge distillation frameworks (Gou et al., 2021; Hinton et al., 2015) especially for  
 86 classification tasks. Formally, we present two versions of our training losses:

87 **Soft Pseudo Labels (SPL):** The predicted class probability distribution. Thus we use Kull-  
 88 back–Leibler divergence (KL-Div) (Kullback and Leibler, 1951) loss function:

$$\mathcal{L}_{\text{KL}} = \sum_{c=1}^{C_m} p_{c,k} \log \left( \frac{p_{c,k}}{q_c} \right) \quad (1)$$

89 where  $p_{c,k}$  is the predicted probability of class  $c$  from the pre-trained model  $\theta_k$  on task  $k$ , and  $q_c$  is  
 90 the predicted probability of class  $c$  from the merged model  $\theta_m$ .

91 **Hard Pseudo Labels (HPL):** The predicted class label in one-hot encoded format. Therefore, the  
 92 cross-entropy loss is applied:

$$\mathcal{L}_{\text{CE}} = - \sum_{c=1}^{C_m} \hat{y}_{c,k} \log(\hat{y}_c) \quad (2)$$

93 Results are shown in 3. We highlight two key observations: (1) Training *StatsMergeLearner* with  
 94 Hard Pseudo Labels (HPL) using cross-entropy loss (KD CE) yields performance comparable to  
 95 training with ground-truth labels (GT CE), achieving 81.2% vs. 88.5% at the task-wise (TW) level  
 96 and 83.5% vs. 90.4% at the layer-wise (LW) level. Importantly, *StatsMerging* eliminates the need for  
 97 manually annotated labels, validating our intuition of leveraging task-specific teacher knowledge for  
 98 supervision. (2) When trained on Soft Pseudo Labels (SPL) using KL-Divergence loss (KL-Div),  
 99 *StatsMergeLearner* underperforms relative to HPL with cross-entropy, obtaining 73.3% vs. 81.2% at  
 100 the TW level and 52.4% vs. 83.5% at the LW level, respectively.

101 We hypothesize that the observed performance drop is due to noisy inter-class relationships within  
 102 the aggregated dataset (Yuan et al., 2021). While a detailed investigation of these relationships is  
 103 beyond the scope of this work on model merging, we believe it presents promising directions for  
 104 future research.

105 Identifies "regularization samples" where soft labels degrade performance due to poor calibration or  
 106 noisy class relationships. Proposes weighted soft labels to mitigate these issues.

Table 3: Multi-task performance (Avg Acc %) of *StatsMerging* when merging ViT-B/32 (4) models on four tasks. *StatsMerging*: shaded in gray. GT: Ground Truth. KD: Knowledge Distillation. GL: Granularity level. TW: Task-wise. LW: Layer-wise.

GL	Loss	CA	EU	RE	GT	Avg Acc
TW	GT CE	73.2	94.2	91.1	95.6	88.5
TW	KD KL-Div	56.5	97.6	56.5	82.4	73.3
TW	KD CE	64.2	88.6	85.2	86.7	81.2
LW	GT CE	75.6	96.3	92.1	97.6	90.4
LW	KD KL-Div	53.1	41.4	65.9	49.1	52.4
LW	KD CE	68.7	91.6	87.2	93.5	83.5

## 107 B.4 Efficient Inference

108 *StatsMergeLearner* is designed to be lightweight, introducing minimal spatial and computational  
 109 overhead to the overall merging process. As shown in Table 4, it contains only 10.99M parameters,  
 110 requires 2.95 GFLOPs, and achieves an inference time of 5.26 ms on an NVIDIA RTX A6000 GPU.

Table 4: Model Size and Computational Overhead of *StatsMergeLearner*

#Params (M)	GFLOPs	Inference Time (ms)
10.99	2.95	5.26

## 111 B.5 Training Curve

The training curve is shown in 1.

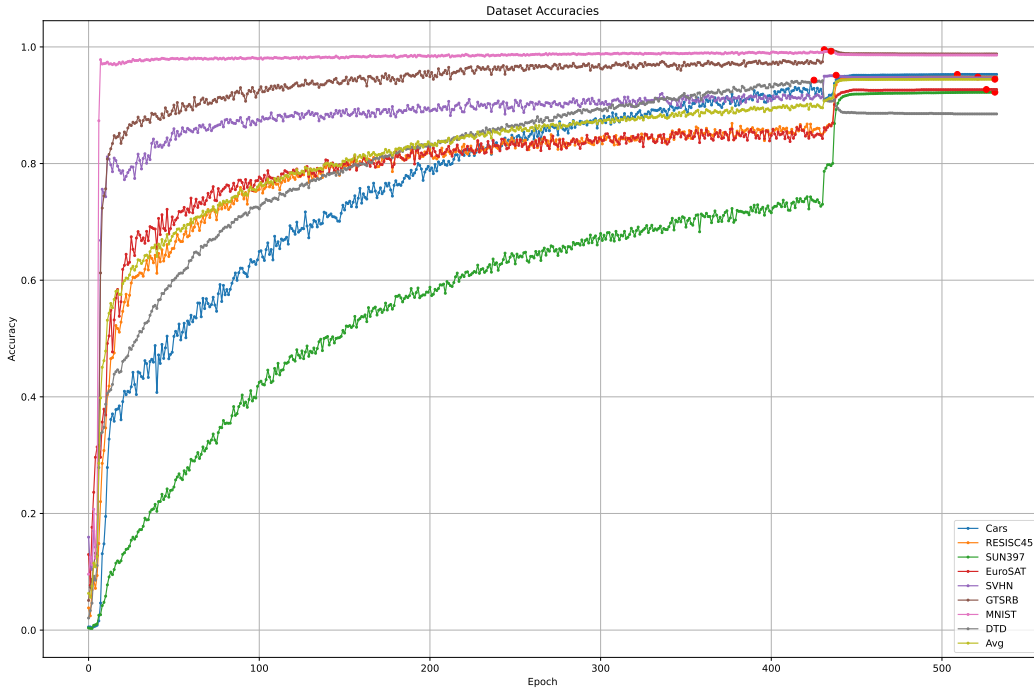


Figure 1: *StatsMerging++* Training Accuracy Curve.

112

## 113 B.6 Extended Related Work

114 **Model Merging Foundations.** Recent efforts in model merging have introduced various strategies  
 115 to efficiently combine multiple models without retraining. Approaches such as ZipIt (Zhang et al.,  
 116 2024a), EMR-Merging (Huang et al., 2024), and Training-Free Pretrained Model Merging methods  
 117 (Sun et al., 2025; Chen et al., 2024) emphasize data-free, tuning-free methodologies, often leveraging  
 118 weight-space heuristics or task-vector alignment. Techniques like Pareto Merging (Chen and Kwok,  
 119 2025), MAP (Li et al., 2024), and  $C^2M^3$  (Crisostomi et al., 2024) formulate model merging as a  
 120 multi-objective or constrained optimization problem to preserve task performance across domains.  
 121 Other works such as Parameter Competition Balancing (Guodong et al.) and Sharpness-Aware  
 122 Fine-Tuning (Lee et al., 2025) address parameter interference during merging. Meanwhile, methods  
 123 like LayerMerge (Kim et al., 2024) and MERGE3 (Mencattini et al., 2025) aim to improve scalability  
 124 and computational efficiency, making merging feasible on consumer-grade hardware.

## B.7 Future Work and Limitations

In this work, we focus on vision-based classification tasks, leaving extensions to other domains, such as object detection, super resolution, and image restoration, for future work. Additionally, expanding this approach to language tasks, particularly large language models (LLMs) (Yang et al., 2024; Song et al., 2024; Zhang et al., 2024b; Tie et al., 2025; Kallini et al., 2025), as well as to multi-modal learning (Zhu et al., 2025; Du et al., 2025; Bousselham et al., 2024; Lin et al., 2024), represents a promising direction for further research.

## References

- Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024.
- I Chen, Hsu-Shen Liu, Wei-Fang Sun, Chen-Hao Chao, Yen-Chang Hsu, Chun-Yi Lee, et al. Retraining-free merging of sparse mixture-of-experts via hierarchical clustering. *arXiv preprint arXiv:2410.08589*, 2024.
- Weiyu Chen and James T. Kwok. Pareto merging: Multi-objective optimization for preference-aware model merging. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL <https://arxiv.org/abs/2408.12105>.
- Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.
- 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.
- Donato Crisostomi, Marco Fumero, Daniele Baieri, Florian Bernard, and Emanuele Rodola.  $c^2m^3$ : Cycle-consistent multi-model merging. *Advances in Neural Information Processing Systems*, 37:28674–28705, 2024.
- Yiyang Du, Xiaochen Wang, Chi Chen, Jiabo Ye, Yiru Wang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Zhifang Sui, et al. Adamms: Model merging for heterogeneous multimodal large language models with unsupervised coefficient optimization. *arXiv preprint arXiv:2503.23733*, 2025.
- Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.
- DU Guodong, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim Kuan Goh, Ho-Kin Tang, Daojing He, et al. Parameter competition balancing for model merging. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- 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.
- Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging. *Advances in Neural Information Processing Systems*, 37:122741–122769, 2024.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023.
- Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *The Eleventh International Conference on Learning Representations*, 2023.

173 Julie Kallini, Shikhar Murty, Christopher D. Manning, Christopher Potts, and Róbert Csordás. Mrt5: Dynamic  
174 token merging for efficient byte-level language models. In *Proceedings of the 13th International Conference on*  
175 *Learning Representations (ICLR 2025)*, 2025. URL <https://openreview.net/forum?id=VYWMq1L7H>.

176 Jinuk Kim, Marwa El Halabi, Mingi Ji, and Hyun Oh Song. Layermerge: neural network depth compression  
177 through layer pruning and merging. *arXiv preprint arXiv:2406.12837*, 2024.

178 Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categoriza-  
179 tion. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages  
180 554–561. IEEE, 2013. ISBN 978-1-4799-3022-7. doi: 10.1109/ICCVW.2013.77.

181 Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical*  
182 *Statistics*, 22(1):79–86, 1951.


183 Yann LeCun, Corinna Cortes, and Christopher JC Burges. The mnist database of handwritten digits.  
184 <http://yann.lecun.com/exdb/mnist/>, 1998.

185 Yeoreum Lee, Jinwook Jung, and Sungyong Baik. Mitigating parameter interference in model merging via  
186 sharpness-aware fine-tuning. *arXiv preprint arXiv:2504.14662*, 2025.

187 Lu Li, Tianyu Zhang, Zhiqi Bu, Suyuchen Wang, Huan He, Jie Fu, Yonghui Wu, Jiang Bian, Yong Chen, and  
188 Yoshua Bengio. Map: Low-compute model merging with amortized pareto fronts via quadratic approximation.  
189 *arXiv preprint arXiv:2406.07529*, 2024. URL <https://arxiv.org/abs/2406.07529>.

190 Ziyi Lin, Dongyang Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Wenqi Shao, Keqin Chen,  
191 Jiaming Han, et al. Sphinx: A mixer of weights, visual embeddings and image scales for multi-modal large  
192 language models. In *European Conference on Computer Vision*, pages 36–55. Springer, 2024.

193 Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural*  
194 *Information Processing Systems*, 35:17703–17716, 2022.

195 Tommaso Mencattini, Adrian Robert Minut, Donato Crisostomi, Andrea Santilli, and Emanuele Rodola. Merge  
196 : Efficient evolutionary merging on consumer-grade gpus. *arXiv preprint arXiv:2502.10436*, 2025.

197 Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in  
198 natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised*  
199 *feature learning*, volume 2011, page 4. Granada, 2011.

200 Woomin Song, Seunghyuk Oh, Sangwoo Mo, Jaehyung Kim, Sukmin Yun, Jung-Woo Ha, and Jinwoo Shin.  
201 Hierarchical context merging: Better long context understanding for pre-trained llms. *arXiv preprint*  
202 *arXiv:2404.10308*, 2024.

203 Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition  
204 benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural*  
205 *networks*, pages 1453–1460. IEEE, 2011.

206 Wenju Sun, Qingyong Li, Yangli-ao Geng, and Boyang Li. Cat merging: A training-free approach for resolving  
207 conflicts in model merging. *arXiv preprint arXiv:2505.06977*, 2025.

208 Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue  
209 Yan, Yao Su, et al. A survey on post-training of large language models. *arXiv preprint arXiv:2503.06072*,  
210 2025.

211 Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a  
212 large collection of scene categories. *International Journal of Computer Vision*, 119:3–22, 2016.

213 Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving  
214 interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115,  
215 2023.

216 Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging:  
217 Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning*  
218 *Representations*.

219 Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model  
220 merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint*  
221 *arXiv:2408.07666*, 2024.



- 222 Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Rethinking soft labels for knowledge  
223 distillation: A bias–variance tradeoff perspective. In *Proceedings of the International Conference on Learning*  
224 *Representations (ICLR)*, 2021. URL [https://openreview.net/forum?id=6x\\_osD4AX9](https://openreview.net/forum?id=6x_osD4AX9).
- 225 Qitian Zhang, Mitchell Wortsman, Simon Kornblith, Rohan Taori, Tatsunori Hashimoto, Benjamin Recht, and  
226 Yair Carmon. Zipit! merging models from different tasks without training. In *International Conference on*  
227 *Learning Representations (ICLR)*, 2024a.
- 228 Yuxin Zhang, Yuxuan Du, Gen Luo, Yunshan Zhong, Zhenyu Zhang, Shiwei Liu, and Rongrong Ji. Cam: Cache  
229 merging for memory-efficient llms inference. In *Forty-first International Conference on Machine Learning*,  
230 2024b.
- 231 Didi Zhu, Yibing Song, Tao Shen, Ziyu Zhao, Jinluan Yang, Min Zhang, and Chao Wu. Remedy: Recipe  
232 merging dynamics in large vision-language models. In *The Thirteenth International Conference on Learning*  
233 *Representations*, 2025.