# Supplementary Material: Task Vector Quantization for Memory-Efficient Model Merging

Youngeun Kim[1]* Seunghwan Lee[2]* Aecheon Jung[2]* Bogon Ryu[2] Sungeun Hong[2]†
[1]Yale University [2]Sungkyunkwan University
youngeun.kim@yale.edu {simon2, kasurashan, bogon.ryu, csehong}@skku.edu

This Appendix provides an overview of our experimental details, further empirical analyses, and additional results. The detailed descriptions of each section are summarized as follows:

- Appendix A: Details of the datasets, model checkpoints, and merging methods used in our experiments.
- Appendix B: Further analyses on the effects of quantization in terms of weight pruning and task vector similarity; includes a sensitivity analysis on bit allocations.
- Appendix C: Detailed results on the 14 and 20 classification tasks, task-level performance, additional quantitative and qualitative results for dense prediction tasks, and loss landscape visualizations.

## A. Experimental Details

### A.1. Employed datasets

We evaluated our approach on a broad and diverse set of datasets that span three major categories: image classification, dense prediction, and natural language processing.

For image classification, we conducted experiments on a total of 19 datasets that cover a wide range of domains and visual characteristics. Specifically, we used SUN397 [41] for scene recognition, Cars [19] for fine-grained vehicle classification, and RESISC45 [3] and EuroSAT [15] for remote sensing imagery. We also included digit and character recognition datasets such as SVHN [24], MNIST [21], EMNIST [7], FashionMNIST [40], and KMNIST [5]. Additional benchmarks included GTSRB [32] for traffic sign recognition, DTD [4] for texture classification, CIFAR-10/100 [20] and STL10 [6] for general object recognition, FER2013 [13] for facial expression recognition, Flowers102 [25] and Oxford-IIIT Pet [26] for fine-grained species classification, PCAM [34] for histopathology image analysis, Food101 [1] for food recognition, and Rendered SST-2 [27], which contains rendered images generated from sentiment classification data.

For dense prediction tasks, we used NYUv2 [30], which provides RGB-D indoor images annotated for multiple tasks, including 13-class semantic segmentation, depth estimation, and surface normal estimation. This dataset was chosen to evaluate the ability of our method to handle multi-modal dense prediction problems.

Finally, for NLP tasks, we adopted the GLUE benchmark [35], which consists of multiple language understanding tasks. We reported the Matthews correlation coefficient for CoLA [38], Pearson and Spearman correlations for STS-B [2], and accuracy for the remaining tasks, including SST-2 [31], MRPC [9], QQP [18], MNLI [39], QNLI [29], and RTE [12]. These benchmarks comprehensively evaluate our method across different input modalities and task types.

### A.2. Model checkpoints

For our experiments, we address three distinct domains. In image classification, we utilize Vision Transformers of varying scales: ViT-B/32 and ViT-L/14, initialized with pretrained CLIP weights [28][1]. To ensure a fair comparison across the 8 classification tasks, we adopt the publicly released model checkpoints from Task Arithmetic [17][2]. For our extended experiments with 14 and 20 classification tasks, we fine-tune CLIP ViT-B/32 following Tall-Mask [36] to obtain task-specific models. For dense prediction tasks, we employ ResNet-50 [14] as the backbone, initializing with ImageNet pre-trained weights. We then fine-tune this model following FusionBench [33] for segmentation, depth estimation, and normal estimation. For NLP tasks, we use RoBERTa-base [22] with the publicly available fine-tuned weights from EMR-Merging [16][3].

### A.3. Merging method baseline

We applied our quantization method against diverse merging strategies, from simple to advanced. Notably, we limited our investigation to approaches that leverage task vectors for the purpose of implementing our proposed quantiza-

*Equal contribution.
†Corresponding author.

[1]https://github.com/mlfoundations/open_clip
[2]https://github.com/mlfoundations/task_vectors
[3]https://github.com/harveyhuang18/EMR_Merging

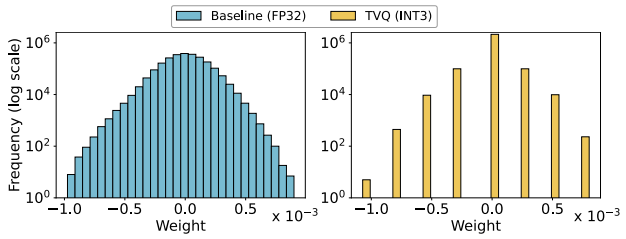



Figure A. Histogram of task vector weight distributions before and after quantization. After quantization, smaller weight values are mapped to zero, leading to a substantial increase in sparsity.

tion scheme. Furthermore, we reimplemented all merging methods and conducted extensive experiments to ensure a fair and comprehensive evaluation. Below, we briefly summarize the key insights of each approach:

**Individual** fine-tunes the pretrained model separately for each task, which optimizes performance for the specific task. However, this approach cannot handle multiple tasks.

**Task Arithmetic** [17] defines a task vector as the difference between a pretrained model and a fine-tuned model, then combines these vectors to form a multi-task model.

**Ties Merging** [42] mitigates interference during merging by addressing redundant parameters and sign conflicts among task vectors.

**MagMax** [23] merges task vectors by selecting, for each parameter, the one with the largest magnitude change.

**Breadcrumbs** [8] applies layer-wise filtering to remove extreme weight changes, including both large outliers and negligible values, constructing a unified multi-task model.

**Consensus TA** [36] retains general weights that are important across multiple tasks while removing selfish weights to reduce task interference.

**LiNeS** [37] applies linear scaling to adjust layer-wise coefficients, which captures the relative importance of each task vector during the merging process.

**AdaMerging** [43] employs an unsupervised approach at test time to optimize merge coefficients, rather than setting them empirically.

**EMR-Merging** [16] constructs a unified model by electing a shared set of weights. It then applies task-specific binary masks and rescaling factors to adjust magnitudes.

**TSV-Compress** [10] exploits the low-rank structure of layer-wise task matrices using singular value decomposition, enabling effective compression of task vectors.

# B. More Empirical Analyses

## B.1. Impact of quantization

In our main experiments, we observed that quantization can occasionally lead to performance improvements. Building on the analysis presented in main paper, which showed that quantization might help prevent overfitting and improve generalization, we further conducted a more extensive investigation using various approaches. In particular, we focused on 3 bit quantization because it exhibited consistent performance gains across diverse tasks.

| Offset \ Base | INT2 | INT3 | INT4 | INT8 |
|---|---|---|---|---|
| INT2 | 69.5 | 70.2 | 69.7 | 69.6 |
| INT3 | **70.9** | 69.9 | 69.5 | 69.5 |
| INT4 | 69.0 | 69.0 | 69.0 | 69.0 |
| INT8 | 68.9 | 69.0 | 69.0 | 69.0 |

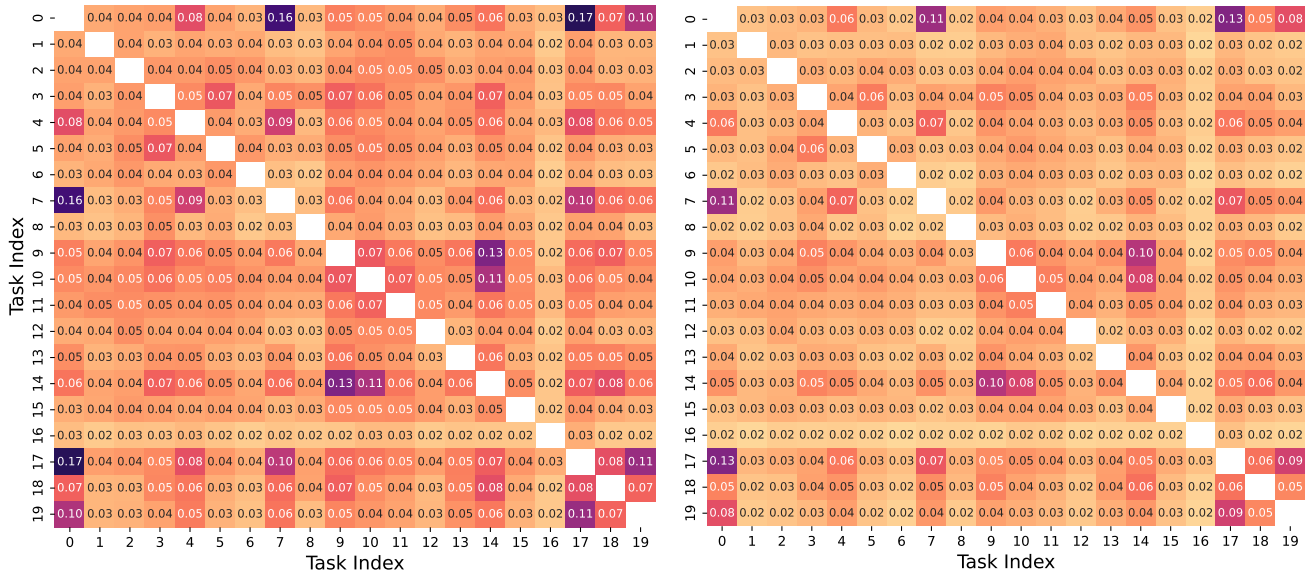
Table A. Average accuracy (%) for various base and offset bitwidth configurations in our Residual Task Vector Quantization (RTVQ). All experiments were performed on merging 8 classification tasks using Task Arithmetic.

**Weight pruning in task vector.** Quantization typically serves as an implicit regularizer, spreading out weight values across discrete levels and thereby reducing overfitting. However, we specifically quantize task vectors, which represent differences between pretrained and fine-tuned model weights, exhibiting a high concentration of values near zero (see Fig.3 in the main paper). To better understand how quantization affects these task vectors, we visualized their weight distributions before and after quantization. As shown in Fig. A, quantization maps small-magnitude weights to exactly zero, effectively pruning less impactful parameters. This process increases the proportion of zero-valued weights to 56.7%, highlighting the pruning effect of quantization in the context of task vectors. Interestingly, this processes mirror strategies commonly employed in task vector-based model merging [8, 16, 36, 42]. Consequently, our quantization approach naturally introduces sparsity, simultaneously providing efficiency gains and potential improvements in model performance.

**Task vector similarity.** We measure cosine similarities of the task vectors before and after quantization by constructing confusion matrices for 20 classification tasks. As shown in Fig. B, quantization reduces off-diagonal similarities, indicating that distinct tasks become more orthogonal. Prior work [17] has observed that task vectors tend to be nearly orthogonal, which facilitates more effective merging. Given that quantization further increases orthogonality, we speculate that this process helps reduce task interference and strengthens the robustness of the merged model.

## B.2. Sensitivity analysis

We conducted a sensitivity analysis of the quantization precision for the base and offset vectors in our Residual Task Vector Quantization (RTVQ). Table A summarizes the performance across 16 distinct bitwidth configurations ranging from 2, 3, 4, 8 bits. While one might expect allocating more bits to either the shared base vector or the task-specific off-




(a) Original task vector (FP32)

(b) Quantized task vector (INT3)

Figure B. Confusion matrices of the cosine similarity among 20 classification task vectors for (a) full-precision (FP32) and (b) 3-bit quantized settings. The diagonal entries are excluded for a clearer comparison. Indices 0 to 19, respectively, correspond to: MNIST, Cars, DTD, EuroSAT, GTSRB, RESISC45, SUN397, SVHN, PCAM, CIFAR100, STL10, OxfordIIITPet, Flowers102, FER2013, CIFAR10, Food101, RenderedSST2, EMNIST, FashionMNIST, and KMNIST.

set vector would yield better performance, our results did not show a clear pattern. Instead, balanced configurations, such as combining a 2-bit base vector with a 3-bit offset vector, provided optimal accuracy. Notably, nearly all configurations surpassed the performance of 2-bit quantization (62.1%) and approached the FP32 baseline (69.2%). These results indicate that RTVQ remains robust across diverse bitwidth configurations, achieving near-FP32 performance even with relatively low-bit allocations.

# C. Additional Results

## C.1. Full results on 14 and 20 classification tasks

In Table B and Table C, we present comprehensive results corresponding to Figure 6 in the main paper, which shows that our quantization method becomes more stable as model size increases. Similarly, stability improves as the number of tasks grows from 14 to 20, particularly for 3-bit TVQ and RTVQ. Additionally, while full-precision models face increasing storage overhead as tasks scale, our quantization method keeps storage requirements manageable. This ensures efficiency gains without compromising accuracy, making it well-suited for large-scale or multi-task scenarios where storage scalability is crucial.

## C.2. Additional results for dense prediction tasks

We provide detailed results corresponding to Table 3 in the main paper for merging dense prediction tasks. Table D expands on these results by comparing a broader range of quantization methods and evaluation metrics. Additionally, Fig. C, D, and E present qualitative visualizations for semantic segmentation, depth estimation, and normal estimation using the state-of-the-art EMR-Merging method. These results further validate the trends consistently observed throughout the paper.

## C.3. Comprehensive task-level results

We report the average accuracy across all tasks in Table 1 and Table 2 of the main paper. This provides an overall comparison of different quantization methods and shows how well each method generalizes across multiple tasks. We also present detailed task-specific results for ViT-B/32 and ViT-L/14 in Table E and F. These results enable closer inspection of per-task behavior and reveal where certain methods perform well or poorly. Such analysis is important for understanding how sensitive quantization strategies are to different task characteristics.

## C.4. Loss landscape visualizations

We utilized the visualization method introduced in [11]. Using 1,024 test images, we computed the loss values across all 16×16 grid points for each task. For 8 vision tasks, we visualized the loss landscapes of all target task and cross-task pairs using ViT-B/32. The visualization results for cross-tasks can be found in Fig. F, G, H, I, while the visualization results for target tasks are presented in Fig. J, K.

| Method | Baseline | FQ | | TVQ (ours) | | | | RTVQ (ours) |
|---|---|---|---|---|---|---|---|---|
| | FP32 | INT8 | INT4 | INT8 | INT4 | INT3 | INT2 | |
| Task arithmetic [17] | 65.4 | 65.2 (-0.2) | 8.7 (-56.7) | 65.4 (0.0) | 65.3 (-0.1) | 65.1 (-0.3) | 60.9 (-4.5) | 65.0 (-0.4) |
| Ties merging [42] | 65.2 | 63.5 (-1.7) | 9.0 (-56.2) | 65.2 (0.0) | 65.0 (-0.2) | 66.4 (1.2) | 61.3 (-3.9) | 63.2 (-2.0) |
| LiNeS [37] | 68.0 | 67.9 (-0.1) | 8.1 (-59.9) | 68.0 (0.0) | 68.0 (0.0) | 67.8 (-0.2) | 62.0 (-6.0) | 67.5 (-0.5) |
| Consensus TA [36] | 70.2 | 64.4 (-5.8) | 8.4 (-61.8) | 70.2 (0.0) | 70.1 (-0.1) | 70.2 (0.0) | 63.2 (-7.0) | 69.8 (-0.4) |
| AdaMerging [43] | 76.7 | 76.2 (-0.5) | 8.2 (-68.5) | 76.7 (0.0) | 76.8 (0.1) | 77.2 (0.5) | 74.4 (-2.3) | 76.1 (-0.6) |
| EMR-Merging [16] | 86.1 | 86.2 (0.1) | 7.7 (-78.4) | 86.3 (0.2) | 88.2 (2.1) | 88.4 (2.3) | 76.2 (-9.9) | 78.9 (-7.2) |

Table B. Comparison of our proposed quantization methods for merging 14 classification tasks using ViT-B/32. Note that our primary objective is to improve storage efficiency for checkpoint saving while minimizing performance degradation relative to the baseline, rather than to maximize performance. The baseline refers to full-precision (FP32) model checkpoints. FQ denotes quantizing fine-tuned checkpoints, TVQ indicates Task Vector Quantization, and RTVQ incorporates our proposed Residual Task Vector Quantization using a 3-bit base vector and a 2-bit offset vector (equivalent to 2.21 bits per task). We report the average accuracy (%) across all tasks, with the difference relative to FP32 shown in parentheses (red indicates a performance drop, green a gain).

| Method | Baseline | FQ | | TVQ (ours) | | | | RTVQ (ours) |
|---|---|---|---|---|---|---|---|---|
| | FP32 | INT8 | INT4 | INT8 | INT4 | INT3 | INT2 | |
| Task arithmetic [17] | 60.8 | 60.5 (-0.3) | 10.4 (-50.4) | 60.8 (0.0) | 60.8 (0.0) | 61.3 (0.5) | 59.5 (-1.3) | 61.0 (0.2) |
| Ties merging [42] | 63.1 | 59.0 (-4.1) | 10.4 (-52.7) | 63.1 (0.0) | 62.9 (-0.2) | 64.1 (1.0) | 60.0 (-3.1) | 58.9 (-4.2) |
| LiNeS [37] | 63.7 | 63.5 (-0.2) | 10.3 (-53.4) | 63.7 (0.0) | 63.7 (0.0) | 64.1 (0.4) | 60.3 (-3.4) | 63.7 (0.0) |
| Consensus TA [36] | 65.0 | 59.3 (-5.7) | 10.5 (-54.5) | 65.0 (0.0) | 65.0 (0.0) | 65.5 (0.5) | 60.7 (-4.3) | 65.7 (0.7) |
| AdaMerging [43] | 69.6 | 69.3 (-0.3) | 9.7 (-59.9) | 71.3 (1.7) | 71.3 (1.7) | 71.8 (2.2) | 70.2 (0.6) | 71.5 (1.9) |
| EMR-Merging [16] | 86.6 | 84.3 (-2.3) | 10.0 (-76.6) | 86.7 (0.1) | 88.4 (1.8) | 87.1 (0.5) | 72.7 (-13.9) | 75.6 (-11.0) |

Table C. Comparison of our proposed quantization methods for merging 20 classification tasks using ViT-B/32. For RTVQ, we use a 3-bit base vector and a 2-bit offset vector (equivalent to 2.15 bits per task)


Figure C. Qualitative results of segmentation task across different quantization methods. RTVQ quantizes both the base vector and offset to 2 bits.

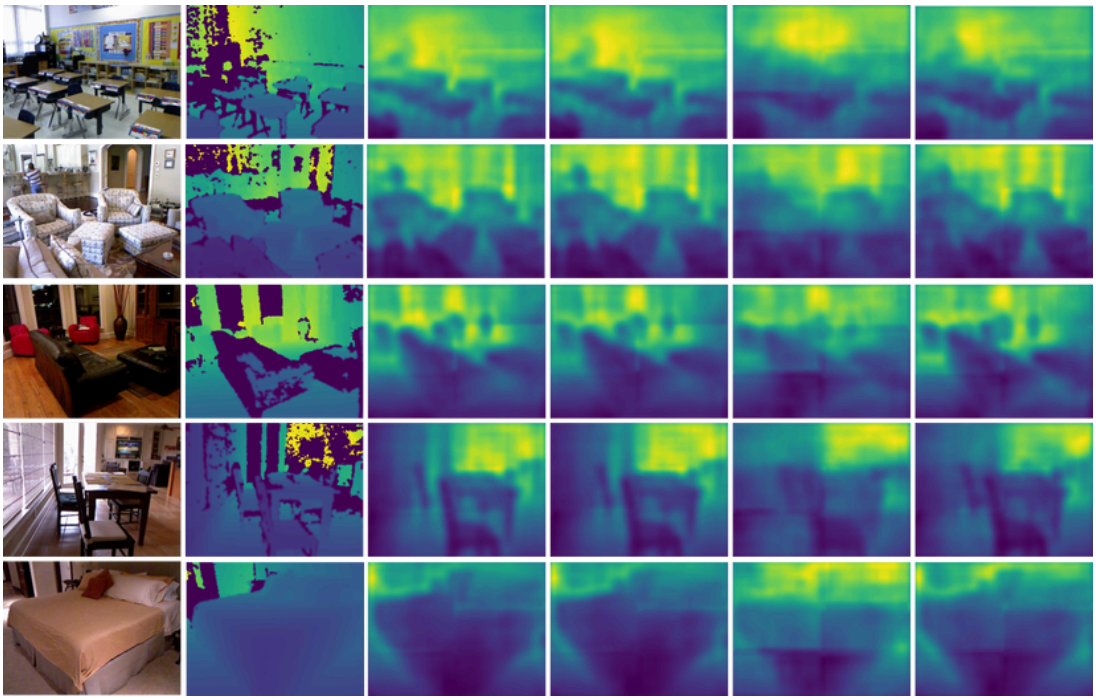



Figure D. Qualitative results of depth estimation task across different quantization methods. RTVQ quantizes both the base vector and offset to 2 bits.

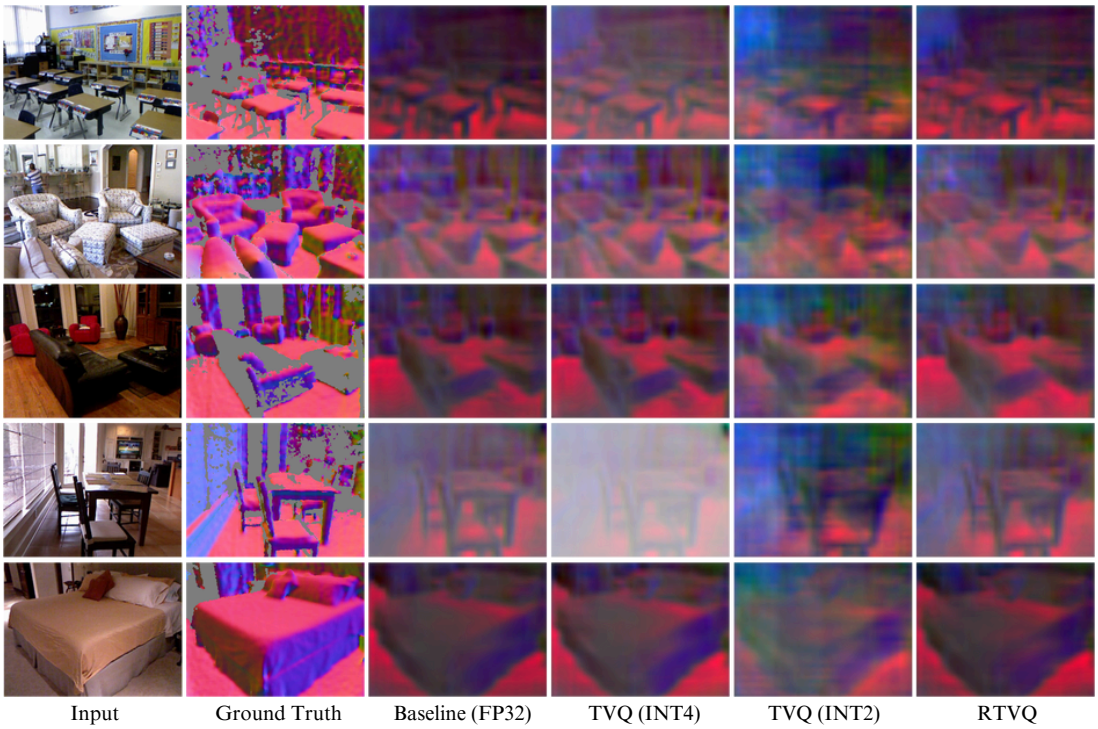



Figure E. Qualitative results of normal estimation task across different quantization methods. RTVQ quantizes both the base vector and offset to 2 bits.

| | | Segmentation | | Depth | | Normal |
|---|---|---|---|---|---|---|
| | | mIoU ↑ | Pix Acc ↑ | Abs Err ↓ | Rel Err ↓ | Mean ↓ |
| **Individual** | FP32 | 52.02 | 74.15 | 41.45 | 17.28 | 24.24 |
| | FQ8 | 51.92 (-0.10) | 73.99 (-0.16) | 41.70 (0.25) | 17.52 (0.24) | 24.34 (0.10) |
| | FQ4 | 1.20 (-50.82) | 15.55 (-58.60) | 89.48 (48.03) | 36.00 (18.72) | 44.24 (20.00) |
| | INT8 | 52.02 (0.00) | 74.15 (0.00) | 41.45 (0.00) | 17.28 (0.00) | 24.24 (0.00) |
| | INT4 | 51.98 (-0.04) | 74.16 (0.01) | 41.44 (-0.01) | 17.35 (0.07) | 24.22 (-0.02) |
| | INT3 | 51.78 (-0.24) | 74.05 (-0.10) | 41.57 (0.12) | 17.35 (0.07) | 24.49 (0.25) |
| | INT2 | 37.67 (-14.35) | 56.25 (-17.90) | 62.46 (21.01) | 24.14 (6.86) | 34.17 (9.93) |
| **Task Arithmetic [17]** | FP32 | 31.60 | 60.31 | 56.67 | 24.03 | 30.62 |
| | FQ8 | 31.68 (0.08) | 60.36 (0.05) | 56.79 (0.12) | 24.01 (-0.02) | 30.63 (0.01) |
| | FQ4 | 6.11 (-25.49) | 23.06 (-37.25) | 79.52 (22.85) | 33.69 (9.66) | 43.67 (13.05) |
| | INT8 | 31.61 (0.01) | 60.32 (0.01) | 56.67 (0.00) | 24.03 (0.00) | 30.62 (0.00) |
| | INT4 | 31.54 (-0.06) | 60.32 (0.01) | 56.65 (-0.02) | 24.04 (0.01) | 30.62 (0.00) |
| | INT3 | 32.11 (0.51) | 60.81 (0.50) | 56.78 (0.11) | 23.98 (-0.05) | 30.72 (0.10) |
| | INT2 | 36.36 (4.76) | 61.32 (1.01) | 63.88 (7.21) | 26.21 (2.18) | 36.07 (5.45) |
| | RTVQ | 36.13 (4.53) | 59.23 (-1.08) | 59.23 (2.56) | 24.63 (0.60) | 32.60 (1.98) |
| **Ties-Merging [42]** | FP32 | 39.91 | 62.70 | 61.25 | 27.28 | 36.17 |
| | FQ8 | 40.14 (0.23) | 63.27 (0.57) | 61.04 (-0.21) | 27.29 (0.01) | 36.19 (0.02) |
| | FQ4 | 9.52 (-30.39) | 25.05 (-37.65) | 77.46 (16.21) | 30.45 (3.17) | 68.94 (32.77) |
| | INT8 | 39.90 (-0.01) | 62.78 (0.08) | 61.21 (-0.04) | 27.26 (-0.02) | 36.16 (-0.01) |
| | INT4 | 39.98 (0.07) | 63.17 (0.47) | 61.04 (-0.21) | 27.21 (-0.07) | 36.20 (0.03) |
| | INT3 | 39.23 (-0.68) | 62.02 (-0.68) | 61.54 (0.29) | 27.26 (-0.02) | 36.37 (0.20) |
| | INT2 | 36.09 (-3.82) | 59.92 (-2.78) | 65.10 (3.85) | 26.54 (-0.74) | 37.03 (0.86) |
| | RTVQ | 37.02 (-2.89) | 63.57 (0.87) | 59.22 (-2.03) | 24.55 (-2.73) | 32.64 (-3.53) |
| **MagMax [23]** | FP32 | 24.73 | 54.71 | 60.27 | 23.88 | 30.25 |
| | FQ8 | 24.80 (0.07) | 54.75 (0.04) | 61.18 (0.91) | 24.17 (0.29) | 30.44 (0.19) |
| | FQ4 | 6.16 (-18.57) | 24.52 (-30.19) | 79.24 (18.97) | 31.60 (7.72) | 39.97 (9.72) |
| | INT8 | 24.76 (0.03) | 54.75 (0.04) | 60.25 (-0.02) | 23.88 (0.00) | 30.26 (0.01) |
| | INT4 | 25.39 (0.66) | 55.19 (0.48) | 61.33 (1.06) | 24.19 (0.31) | 30.04 (-0.21) |
| | INT3 | 23.28 (-1.45) | 52.88 (-1.83) | 62.87 (2.60) | 24.59 (0.71) | 29.77 (-0.48) |
| | INT2 | 29.93 (5.20) | 58.65 (3.94) | 64.29 (4.02) | 25.59 (1.71) | 32.22 (1.97) |
| | RTVQ | 29.39 (4.66) | 58.48 (3.77) | 62.57 (2.30) | 24.73 (0.85) | 31.06 (0.81) |
| **Breadcrumbs [8]** | FP32 | 34.14 | 58.51 | 66.05 | 27.17 | 36.85 |
| | FQ8 | 34.29 (0.15) | 58.80 (0.29) | 65.97 (-0.08) | 27.15 (-0.02) | 36.84 (-0.01) |
| | FQ4 | 19.73 (-14.41) | 40.63 (-17.88) | 75.86 (9.81) | 29.45 (2.28) | 40.60 (3.75) |
| | INT8 | 34.19 (0.05) | 58.57 (0.06) | 66.06 (0.01) | 27.17 (0.00) | 36.88 (0.03) |
| | INT4 | 34.26 (0.12) | 58.56 (0.05) | 66.07 (0.02) | 27.18 (0.01) | 37.00 (0.15) |
| | INT3 | 34.30 (0.16) | 58.70 (0.19) | 66.11 (0.06) | 27.15 (-0.02) | 36.86 (0.01) |
| | INT2 | 32.19 (-1.95) | 54.22 (-4.29) | 69.03 (2.98) | 28.44 (1.27) | 40.58 (3.73) |
| | RTVQ | 33.97 (-0.17) | 57.31 (-1.20) | 67.13 (1.08) | 27.66 (0.49) | 38.29 (1.44) |
| **EMR-Merging [16]** | FP32 | 41.50 | 67.24 | 48.59 | 19.44 | 26.52 |
| | FQ8 | 41.73 (0.23) | 67.27 (0.03) | 48.64 (0.05) | 19.37 (-0.07) | 26.54 (0.02) |
| | FQ4 | 1.20 (-40.30) | 15.55 (-51.69) | 80.46 (31.87) | 35.39 (15.95) | 45.78 (19.26) |
| | INT8 | 41.69 (0.19) | 67.37 (0.13) | 48.43 (-0.16) | 19.35 (-0.09) | 26.50 (-0.02) |
| | INT4 | 44.79 (3.29) | 69.53 (2.29) | 47.13 (-1.46) | 18.76 (-0.68) | 26.56 (0.04) |
| | INT3 | 43.00 (1.50) | 67.66 (0.42) | 48.05 (-0.54) | 18.87 (-0.57) | 29.21 (2.69) |
| | INT2 | 21.33 (-20.17) | 40.95 (-26.29) | 68.31 (19.72) | 25.45 (6.01) | 45.16 (18.64) |
| | RTVQ | 34.11 (-7.39) | 57.23 (-10.01) | 59.31 (10.72) | 22.07 (2.63) | 34.99 (8.47) |

Table D. Comprehensive experimental results of merging ResNet-50 models on three NYUv2 tasks. We compare the proposed quantization methods against the full-precision baseline (FP32). RTVQ quantizes both the base vector and offset vector to 2 bits.

| Method | | SUN | Cars | RES. | Euro | SVH. | GTS. | MNI. | DTD | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| **Individual** | FP32 | 75.3 | 77.7 | 96.1 | 99.7 | 97.5 | 98.7 | 99.7 | 79.4 | 90.5 |
| | FQ8 | 74.9 (-0.4) | 77.2 (-0.5) | 96.0 (-0.1) | 99.9 (0.2) | 97.4 (-0.1) | 98.7 (0.0) | 99.7 (0.0) | 79.7 (0.3) | 90.4 (-0.1) |
| | FQ4 | 0.3 (-75.0) | 0.7 (-77.0) | 3.2 (-92.9) | 7.5 (-92.2) | 8.9 (-88.6) | 2.2 (-96.5) | 8.7 (-91.0) | 2.0 (-77.4) | 4.2 (-86.3) |
| | INT8 | 75.3 (0.0) | 77.7 (0.0) | 96.1 (0.0) | 99.9 (0.2) | 97.5 (0.0) | 98.7 (0.0) | 99.7 (0.0) | 79.4 (0.0) | 90.5 (0.0) |
| | INT4 | 75.2 (-0.1) | 77.8 (0.1) | 96.1 (0.0) | 99.9 (0.2) | 97.4 (-0.1) | 98.7 (0.0) | 99.7 (0.0) | 79.4 (0.0) | 90.5 (0.0) |
| | INT3 | 75.4 (0.1) | 78.9 (1.2) | 96.0 (-0.1) | 99.9 (0.2) | 97.3 (-0.2) | 99.0 (0.3) | 99.7 (0.0) | 79.4 (0.0) | 90.7 (0.2) |
| | INT2 | 73.8 (-1.5) | 73.3 (-4.4) | 91.2 (-4.9) | 94.8 (-4.9) | 81.4 (-16.1) | 81.8 (-16.9) | 98.5 (-1.2) | 73.2 (-6.2) | 83.5 (-7.0) |
| **Task Arithmetic [17]** | FP32 | 55.2 | 54.9 | 66.7 | 78.9 | 80.2 | 69.7 | 97.3 | 50.4 | 69.2 |
| | FQ8 | 53.7 (-1.5) | 52.7 (-2.2) | 66.2 (-0.5) | 75.5 (-3.4) | 79.7 (-0.5) | 69.0 (-0.7) | 97.2 (-0.1) | 50.6 (0.2) | 68.1 (-1.1) |
| | FQ4 | 0.3 (-54.9) | 0.6 (-54.3) | 1.7 (-65.0) | 7.0 (-71.9) | 9.1 (-71.1) | 3.8 (-65.9) | 8.9 (-88.4) | 2.3 (-48.1) | 4.2 (-65.0) |
| | INT8 | 55.2 (0.0) | 55.0 (0.1) | 66.7 (0.0) | 77.4 (-1.5) | 80.2 (0.0) | 69.7 (0.0) | 97.3 (0.0) | 50.1 (-0.3) | 69.0 (-0.2) |
| | INT4 | 55.6 (0.4) | 55.0 (0.1) | 66.9 (0.2) | 77.7 (-1.2) | 80.2 (0.0) | 69.6 (-0.1) | 97.3 (0.0) | 50.5 (0.1) | 69.1 (-0.1) |
| | INT3 | 60.8 (5.6) | 59.0 (4.1) | 71.6 (4.9) | 79.5 (0.6) | 79.2 (-1.0) | 69.1 (-0.6) | 97.1 (-0.2) | 53.5 (3.1) | 71.2 (2.0) |
| | INT2 | 66.1 (10.9) | 62.4 (7.5) | 69.8 (3.1) | 66.6 (-12.3) | 54.6 (-25.6) | 43.2 (-26.5) | 83.0 (-14.3) | 50.8 (0.4) | 62.1 (-7.1) |
| | RTVQ | 58.1 (2.9) | 56.8 (1.9) | 70.4 (3.7) | 81.8 (2.9) | 77.6 (-2.6) | 67.6 (-2.1) | 96.8 (-0.5) | 52.9 (2.5) | 70.2 (1.0) |
| **Ties-Merging [42]** | FP32 | 65.0 | 64.4 | 74.8 | 77.4 | 81.2 | 69.3 | 96.5 | 54.5 | 72.9 |
| | FQ8 | 56.1 (-8.9) | 52.6 (-11.8) | 63.8 (-11.0) | 67.9 (-9.5) | 73.6 (-7.6) | 62.3 (-7.0) | 94.7 (-1.8) | 47.1 (-7.4) | 64.8 (-8.1) |
| | FQ4 | 0.2 (-64.8) | 0.5 (-63.9) | 1.9 (-72.9) | 9.0 (-68.4) | 7.5 (-73.7) | 2.2 (-67.1) | 9.6 (-86.9) | 2.4 (-52.1) | 4.2 (-68.7) |
| | INT8 | 65.0 (0.0) | 64.2 (-0.2) | 74.6 (-0.2) | 76.7 (-0.7) | 81.2 (0.0) | 69.4 (0.1) | 96.6 (0.1) | 54.2 (-0.3) | 72.7 (-0.2) |
| | INT4 | 63.8 (-1.2) | 63.5 (-0.9) | 74.6 (-0.2) | 77.5 (0.1) | 78.9 (-2.3) | 69.1 (-0.2) | 95.7 (-0.8) | 52.9 (-1.6) | 72.0 (-0.9) |
| | INT3 | 64.5 (-0.5) | 64.1 (-0.3) | 71.9 (-2.9) | 74.0 (-3.4) | 88.6 (7.4) | 74.0 (4.7) | 98.2 (1.7) | 53.4 (-1.1) | 73.6 (0.7) |
| | INT2 | 66.3 (1.3) | 62.9 (-1.5) | 69.7 (-5.1) | 66.4 (-11.0) | 56.4 (-24.8) | 43.7 (-25.6) | 84.8 (-11.7) | 50.6 (-3.9) | 62.6 (-10.3) |
| | RTVQ | 61.9 (-3.1) | 64.1 (-0.3) | 77.3 (2.5) | 75.9 (-1.5) | 81.8 (0.6) | 68.6 (-0.7) | 96.4 (-0.1) | 55.9 (1.4) | 72.7 (-0.2) |
| **LiNes [37]** | FP32 | 63.7 | 63.9 | 75.1 | 86.1 | 79.4 | 72.2 | 96.2 | 56.5 | 74.1 |
| | FQ8 | 63.6 (-0.1) | 62.7 (-1.2) | 75.1 (0.0) | 85.6 (-0.5) | 79.4 (0.0) | 72.6 (0.4) | 96.0 (-0.2) | 56.3 (-0.2) | 73.9 (-0.2) |
| | FQ4 | 0.2 (-63.5) | 0.3 (-63.6) | 2.5 (-72.6) | 8.9 (-77.2) | 7.7 (-71.7) | 5.9 (-66.3) | 6.9 (-89.3) | 1.7 (-54.8) | 4.3 (-69.8) |
| | INT8 | 63.7 (0.0) | 63.9 (0.0) | 75.1 (0.0) | 86.2 (0.1) | 79.4 (0.0) | 72.2 (0.0) | 96.2 (0.0) | 56.5 (0.0) | 74.2 (0.1) |
| | INT4 | 63.9 (0.2) | 64.2 (0.3) | 75.5 (0.4) | 86.1 (0.0) | 79.3 (-0.1) | 72.1 (-0.1) | 96.1 (-0.1) | 56.8 (0.3) | 74.2 (0.1) |
| | INT3 | 66.6 (2.9) | 65.9 (2.0) | 77.7 (2.6) | 85.7 (-0.4) | 77.2 (-2.2) | 70.6 (-1.6) | 95.7 (-0.5) | 58.0 (1.5) | 74.7 (0.6) |
| | INT2 | 66.4 (2.7) | 63.1 (-0.8) | 71.0 (-4.1) | 64.9 (-21.2) | 47.5 (-31.9) | 42.7 (-29.5) | 78.8 (-17.4) | 51.2 (-5.3) | 60.7 (-13.4) |
| | RTVQ | 65.4 (1.7) | 64.3 (0.4) | 77.6 (2.5) | 86.7 (0.6) | 75.8 (-3.6) | 69.5 (-2.7) | 95.3 (-0.9) | 58.7 (2.2) | 74.2 (0.1) |
| **Consensus TA [36]** | FP32 | 64.8 | 63.1 | 72.2 | 82.6 | 84.4 | 77.2 | 97.2 | 57.8 | 74.9 |
| | FQ8 | 61.0 (-3.8) | 52.7 (-10.4) | 68.9 (-3.3) | 77.3 (-5.3) | 82.5 (-1.9) | 73.1 (-4.1) | 93.6 (-3.6) | 55.3 (-2.5) | 70.6 (-4.3) |
| | FQ4 | 0.4 (-64.4) | 0.5 (-62.6) | 1.8 (-70.4) | 6.8 (-75.8) | 8.0 (-76.4) | 1.9 (-75.3) | 8.2 (-89.0) | 2.1 (-55.7) | 3.7 (-71.2) |
| | INT8 | 64.8 (0.0) | 63.1 (0.0) | 72.2 (0.0) | 82.5 (-0.1) | 84.4 (0.0) | 77.2 (0.0) | 97.3 (0.1) | 57.7 (-0.1) | 74.9 (0.0) |
| | INT4 | 64.9 (0.1) | 63.1 (0.0) | 72.7 (0.5) | 82.8 (0.2) | 84.2 (-0.2) | 76.7 (-0.5) | 97.2 (0.0) | 57.8 (0.0) | 74.9 (0.0) |
| | INT3 | 66.5 (1.7) | 64.6 (1.5) | 75.0 (2.8) | 81.8 (-0.8) | 81.4 (-3.0) | 72.5 (-4.7) | 96.6 (-0.6) | 59.8 (2.0) | 74.8 (-0.1) |
| | INT2 | 66.5 (1.7) | 61.2 (-1.9) | 69.5 (-2.7) | 61.9 (-20.7) | 46.4 (-38.0) | 37.8 (-39.4) | 71.0 (-26.2) | 53.4 (-4.4) | 58.5 (-16.4) |
| | RTVQ | 67.7 (2.9) | 63.9 (0.8) | 76.6 (4.4) | 79.9 (-2.7) | 73.1 (-11.3) | 64.7 (-12.5) | 94.3 (-2.9) | 61.2 (3.4) | 72.7 (-2.2) |
| **AdaMerging [43]** | FP32 | 64.7 | 70.2 | 83.6 | 94.2 | 85.7 | 94.2 | 97.7 | 63.7 | 81.8 |
| | FQ8 | 63.5 (-1.2) | 69.8 (-0.4) | 83.9 (0.3) | 93.1 (-1.1) | 87.2 (1.5) | 94.4 (0.2) | 97.9 (0.2) | 63.1 (-0.6) | 81.6 (-0.2) |
| | FQ4 | 0.3 (-64.4) | 0.5 (-69.7) | 2.7 (-80.9) | 9.0 (-85.2) | 8.8 (-76.9) | 4.0 (-90.2) | 8.9 (-88.8) | 2.1 (-61.6) | 4.5 (-77.3) |
| | INT8 | 64.4 (-0.3) | 70.5 (0.3) | 83.7 (0.1) | 92.7 (-1.5) | 86.8 (1.1) | 94.3 (0.1) | 97.5 (-0.2) | 62.6 (-1.1) | 81.6 (-0.2) |
| | INT4 | 64.1 (-0.6) | 70.4 (0.2) | 83.2 (-0.4) | 93.3 (-0.9) | 86.7 (1.0) | 94.2 (0.0) | 97.7 (0.0) | 62.1 (-1.6) | 81.5 (-0.3) |
| | INT3 | 64.3 (-0.4) | 71.3 (1.1) | 83.8 (0.2) | 93.0 (-1.2) | 87.9 (2.2) | 94.3 (0.1) | 97.8 (0.1) | 63.5 (-0.2) | 82.0 (0.2) |
| | INT2 | 63.4 (-1.3) | 64.8 (-5.4) | 84.0 (0.4) | 90.7 (-3.5) | 88.5 (2.8) | 71.2 (-23.0) | 98.5 (0.8) | 63.7 (0.0) | 78.1 (-3.7) |
| | RTVQ | 66.3 (1.6) | 71.5 (1.3) | 84.6 (1.0) | 92.7 (-1.5) | 88.2 (2.5) | 93.5 (-0.7) | 98.1 (0.4) | 67.8 (4.1) | 82.8 (1.0) |
| **EMR-Merging [16]** | FP32 | 71.8 | 72.8 | 93.5 | 99.4 | 96.9 | 98.1 | 99.6 | 74.4 | 88.3 |
| | FQ8 | 73.6 (1.8) | 75.9 (3.1) | 94.0 (0.5) | 99.0 (-0.4) | 96.6 (-0.3) | 97.2 (-0.9) | 99.6 (0.0) | 73.4 (-1.0) | 88.7 (0.4) |
| | FQ4 | 0.3 (-71.5) | 0.6 (-72.2) | 2.8 (-90.7) | 5.8 (-93.6) | 9.5 (-87.4) | 1.5 (-96.6) | 8.0 (-91.6) | 2.7 (-71.7) | 3.9 (-84.4) |
| | INT8 | 72.0 (0.2) | 72.9 (0.1) | 93.7 (0.2) | 99.4 (0.0) | 96.9 (0.0) | 98.2 (0.1) | 99.6 (0.0) | 74.5 (0.1) | 88.4 (0.1) |
| | INT4 | 74.1 (2.3) | 76.4 (3.6) | 94.9 (1.4) | 99.7 (0.3) | 97.3 (0.4) | 98.6 (0.5) | 99.6 (0.0) | 77.4 (3.0) | 89.8 (1.5) |
| | INT3 | 74.9 (3.1) | 78.0 (5.2) | 95.2 (1.7) | 99.6 (0.2) | 97.2 (0.3) | 98.4 (0.3) | 99.7 (0.1) | 77.1 (2.7) | 90.0 (1.7) |
| | INT2 | 71.5 (-0.3) | 70.5 (-2.3) | 84.9 (-8.6) | 89.0 (-10.4) | 72.4 (-24.5) | 68.0 (-30.1) | 96.3 (-3.3) | 64.9 (-9.5) | 77.2 (-11.1) |
| | RTVQ | 71.4 (-0.4) | 71.4 (-1.4) | 88.6 (-4.9) | 94.2 (-5.2) | 89.7 (-7.2) | 84.3 (-13.8) | 99.0 (-0.6) | 66.6 (-7.8) | 83.2 (-5.1) |

Table E. Comprehensive task-level results for merging 8 classification tasks using ViT-B/32. For RTVQ, we use a 3-bit base vector and a 2-bit offset vector (equivalent to 2.375 bits per task).

| Method | | SUN | Cars | RES. | Euro | SVH. | GTS. | MNI. | DTD | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| **Individual** | FP32 | 82.3 | 92.4 | 97.4 | 100 | 98.1 | 99.2 | 99.7 | 84.1 | 94.2 |
| | FT8 | 82.3 (0.0) | 92.2 (-0.2) | 97.3 (-0.1) | 99.9 (-0.1) | 98.1 (0.0) | 99.2 (0.0) | 99.7 (0.0) | 84.7 (0.6) | 94.2 (0.0) |
| | FT4 | 0.2 (-82.1) | 0.5 (-91.9) | 2.7 (-94.7) | 4.0 (-96.0) | 8.0 (-90.1) | 2.4 (-96.8) | 14.8 (-84.9) | 1.8 (-82.3) | 4.3 (-89.9) |
| | INT8 | 82.3 (0.0) | 92.4 (0.0) | 97.4 (0.0) | 99.9 (-0.1) | 98.1 (0.0) | 99.2 (0.0) | 99.7 (0.0) | 84.1 (0.0) | 94.1 (-0.1) |
| | INT4 | 82.3 (0.0) | 92.4 (0.0) | 97.4 (0.0) | 99.9 (-0.1) | 98.1 (0.0) | 99.2 (0.0) | 99.7 (0.0) | 84.3 (0.2) | 94.2 (0.0) |
| | INT3 | 82.3 (0.0) | 92.5 (0.1) | 97.4 (0.0) | 99.9 (-0.1) | 98.1 (0.0) | 99.2 (0.0) | 99.7 (0.0) | 84.3 (0.2) | 94.2 (0.0) |
| | INT2 | 80.5 (-1.8) | 90.9 (-1.5) | 96.3 (-1.1) | 99.3 (-0.7) | 87.5 (-10.6) | 93.7 (-5.5) | 99.3 (-0.4) | 80.0 (-4.1) | 90.9 (-3.3) |
| **Task Arithmetic [17]** | FP32 | 74.1 | 82.1 | 86.7 | 92.6 | 87.9 | 86.8 | 98.9 | 65.6 | 84.3 |
| | FQ8 | 73.8 (-0.3) | 81.6 (-0.5) | 86.3 (-0.4) | 92.9 (0.3) | 87.6 (-0.3) | 86.1 (-0.7) | 98.9 (0.0) | 65.6 (0.0) | 84.1 (-0.2) |
| | FQ4 | 0.2 (-73.9) | 0.5 (-81.6) | 1.5 (-85.2) | 9.6 (-83.0) | 9.7 (-78.2) | 2.1 (-84.7) | 9.8 (-89.1) | 2.4 (-63.2) | 4.5 (-79.8) |
| | INT8 | 74.1 (0.0) | 82.1 (0.0) | 86.7 (0.0) | 92.6 (0.0) | 87.9 (0.0) | 86.8 (0.0) | 98.9 (0.0) | 65.6 (0.0) | 84.3 (0.0) |
| | INT4 | 74.1 (0.0) | 82.2 (0.1) | 86.7 (0.0) | 92.7 (0.1) | 87.8 (-0.1) | 86.7 (-0.1) | 98.9 (0.0) | 65.9 (0.3) | 84.4 (0.1) |
| | INT3 | 74.5 (0.4) | 83.3 (1.2) | 87.7 (1.0) | 93.5 (0.9) | 87.0 (-0.9) | 86.9 (0.1) | 98.9 (0.0) | 66.9 (1.3) | 84.8 (0.5) |
| | INT2 | 72.8 (-1.3) | 82.8 (0.7) | 84.0 (-2.7) | 86.1 (-6.5) | 72.9 (-15.0) | 65.5 (-21.3) | 94.2 (-4.7) | 64.7 (-0.9) | 77.9 (-6.4) |
| | RTVQ | 75.0 (0.9) | 83.0 (0.9) | 87.8 (1.1) | 93.0 (0.4) | 86.0 (-1.9) | 86.6 (-0.2) | 98.8 (-0.1) | 68.0 (2.4) | 84.8 (0.5) |
| **Ties-Merging [42]** | FP32 | 75.0 | 84.5 | 88.0 | 94.3 | 85.7 | 82.1 | 98.7 | 67.7 | 84.5 |
| | FQ8 | 71.2 (-3.8) | 77.0 (-7.5) | 79.5 (-8.5) | 86.0 (-8.3) | 80.8 (-4.9) | 72.8 (-9.3) | 97.9 (-0.8) | 61.5 (-6.2) | 78.3 (-6.2) |
| | FQ4 | 0.2 (-74.8) | 0.5 (-84.0) | 4.5 (-83.5) | 5.1 (-89.2) | 7.8 (-77.9) | 3.0 (-79.1) | 10.3 (-88.4) | 1.8 (-65.9) | 4.2 (-80.3) |
| | INT8 | 75.0 (0.0) | 84.5 (0.0) | 88.0 (0.0) | 94.3 (0.0) | 85.7 (0.0) | 82.1 (0.0) | 98.7 (0.0) | 67.8 (0.1) | 84.5 (0.0) |
| | INT4 | 75.0 (0.0) | 84.4 (-0.1) | 88.2 (0.2) | 94.3 (0.0) | 85.9 (0.2) | 82.9 (0.8) | 98.6 (-0.1) | 67.8 (0.1) | 84.6 (0.1) |
| | INT3 | 76.4 (1.4) | 85.7 (1.2) | 88.0 (0.0) | 93.6 (-0.7) | 89.0 (3.3) | 82.6 (0.5) | 99.1 (0.4) | 67.7 (0.0) | 85.3 (0.8) |
| | INT2 | 73.3 (-1.7) | 83.1 (-1.4) | 84.4 (-3.6) | 85.9 (-8.4) | 73.8 (-11.9) | 64.2 (-17.9) | 94.8 (-3.9) | 64.5 (-3.2) | 78.0 (-6.5) |
| | RTVQ | 74.9 (-0.1) | 83.6 (-0.9) | 86.1 (-1.9) | 91.5 (-2.8) | 79.7 (-6.0) | 72.6 (-9.5) | 97.6 (-1.1) | 67.1 (-0.6) | 81.6 (-2.9) |
| **LiNes [37]** | FP32 | 74.5 | 85.4 | 88.8 | 95.4 | 90.8 | 90.8 | 99.3 | 70.4 | 86.9 |
| | FQ8 | 73.9 (-0.6) | 84.2 (-1.2) | 88.2 (-0.6) | 95.4 (0.0) | 90.3 (-0.5) | 89.8 (-1.0) | 99.3 (0.0) | 69.7 (-0.7) | 86.4 (-0.5) |
| | FQ4 | 0.3 (-74.2) | 0.5 (-84.9) | 3.6 (-85.2) | 13.7 (-81.7) | 8.8 (-82.0) | 4.3 (-86.5) | 9.8 (-89.5) | 2.2 (-68.2) | 5.4 (-81.5) |
| | INT8 | 74.5 (0.0) | 85.4 (0.0) | 88.8 (0.0) | 95.4 (0.0) | 90.8 (0.0) | 90.8 (0.0) | 99.3 (0.0) | 70.4 (0.0) | 86.9 (0.0) |
| | INT4 | 74.5 (0.0) | 85.4 (0.0) | 88.8 (0.0) | 95.4 (0.0) | 90.7 (-0.1) | 90.8 (0.0) | 99.3 (0.0) | 70.6 (0.2) | 86.9 (0.0) |
| | INT3 | 75.4 (0.9) | 86.5 (1.1) | 90.4 (1.6) | 95.8 (0.4) | 90.2 (-0.6) | 91.6 (0.8) | 99.3 (0.0) | 72.4 (2.0) | 87.7 (0.8) |
| | INT2 | 75.4 (0.9) | 85.8 (0.4) | 88.7 (-0.1) | 91.4 (-4.0) | 74.6 (-16.2) | 72.7 (-18.1) | 95.7 (-3.6) | 69.7 (-0.7) | 81.8 (-5.1) |
| | RTVQ | 75.3 (0.8) | 86.4 (1.0) | 90.2 (1.4) | 95.4 (0.0) | 89.3 (-1.5) | 91.3 (0.5) | 99.3 (0.0) | 74.1 (3.7) | 87.7 (0.8) |
| **Consensus TA [36]** | FP32 | 74.9 | 83.0 | 88.1 | 95.4 | 91.3 | 91.5 | 99.1 | 69.6 | 86.6 |
| | FQ8 | 73.0 (-1.9) | 78.7 (-4.3) | 85.5 (-2.6) | 94.7 (-0.7) | 89.4 (-1.9) | 89.2 (-2.3) | 98.8 (-0.3) | 67.5 (-2.1) | 84.6 (-2.0) |
| | FQ4 | 0.1 (-74.8) | 0.5 (-82.5) | 3.2 (-84.9) | 9.2 (-86.2) | 6.7 (-84.6) | 2.1 (-89.4) | 9.8 (-89.3) | 1.6 (-68.0) | 4.2 (-82.4) |
| | INT8 | 74.9 (0.0) | 83.0 (0.0) | 88.1 (0.0) | 95.4 (0.0) | 91.3 (0.0) | 91.5 (0.0) | 99.1 (0.0) | 69.5 (-0.1) | 86.6 (0.0) |
| | INT4 | 75.0 (0.1) | 82.8 (-0.2) | 88.2 (0.1) | 95.5 (0.1) | 91.2 (-0.1) | 91.5 (0.0) | 99.1 (0.0) | 69.8 (0.2) | 86.6 (0.0) |
| | INT3 | 75.4 (0.5) | 83.9 (0.9) | 89.5 (1.4) | 95.7 (0.3) | 89.9 (-1.4) | 91.7 (0.2) | 99.0 (-0.1) | 71.7 (2.1) | 87.1 (0.5) |
| | INT2 | 74.3 (-0.6) | 83.8 (0.8) | 86.9 (-1.2) | 85.3 (-10.1) | 70.7 (-20.6) | 69.0 (-22.5) | 92.0 (-7.1) | 70.1 (0.5) | 79.0 (-7.6) |
| | RTVQ | 76.3 (1.4) | 84.8 (1.8) | 90.5 (2.4) | 95.4 (0.0) | 87.1 (-4.2) | 90.9 (-0.6) | 98.6 (-0.5) | 75.0 (5.4) | 87.3 (0.7) |
| **AdaMerging [43]** | FP32 | 77.0 | 90.6 | 91.2 | 96.6 | 93.5 | 97.8 | 99.1 | 80.2 | 90.8 |
| | FQ8 | 76.9 (-0.1) | 90.5 (-0.1) | 91.5 (0.3) | 96.2 (-0.4) | 93.6 (0.1) | 97.9 (0.1) | 99.0 (-0.1) | 80.5 (0.3) | 90.8 (0.0) |
| | FQ4 | 0.3 (-76.7) | 0.5 (-90.1) | 2.1 (-89.1) | 11.7 (-84.9) | 9.7 (-83.8) | 2.1 (-95.7) | 9.8 (-89.3) | 2.1 (-78.1) | 4.8 (-86.0) |
| | INT8 | 77.3 (0.3) | 90.7 (0.1) | 91.5 (0.3) | 96.6 (0.0) | 93.9 (0.4) | 98.1 (0.3) | 99.0 (-0.1) | 80.2 (0.0) | 90.9 (0.1) |
| | INT4 | 77.2 (0.2) | 90.7 (0.1) | 91.9 (0.7) | 96.1 (-0.5) | 93.8 (0.3) | 98.0 (0.2) | 99.1 (0.0) | 80.3 (0.1) | 90.9 (0.1) |
| | INT3 | 77.3 (0.3) | 90.7 (0.1) | 91.8 (0.6) | 96.1 (-0.5) | 93.7 (0.2) | 98.3 (0.5) | 99.1 (0.0) | 80.7 (0.5) | 91.0 (0.2) |
| | INT2 | 77.7 (0.7) | 88.8 (-1.8) | 92.3 (1.1) | 96.4 (-0.2) | 88.6 (-4.9) | 94.6 (-3.2) | 99.0 (-0.1) | 78.1 (-2.1) | 89.4 (-1.4) |
| | RTVQ | 76.8 (-0.2) | 90.7 (0.1) | 91.7 (0.5) | 96.6 (0.0) | 92.9 (-0.6) | 97.9 (0.1) | 99.2 (0.1) | 81.2 (1.0) | 90.9 (0.1) |
| **EMR-Merging [16]** | FP32 | 81.1 | 90.7 | 96.8 | 99.7 | 97.9 | 99.1 | 99.7 | 82.7 | 93.5 |
| | FQ8 | 80.3 (-0.8) | 90.8 (0.1) | 96.4 (-0.4) | 99.7 (0.0) | 97.3 (-0.6) | 98.2 (-0.9) | 99.6 (-0.1) | 79.8 (-2.9) | 92.8 (-0.7) |
| | FQ4 | 0.2 (-80.9) | 0.6 (-90.1) | 2.5 (-94.3) | 5.1 (-94.6) | 8.2 (-89.7) | 2.7 (-96.4) | 13.9 (-85.8) | 2.3 (-80.4) | 4.4 (-89.1) |
| | INT8 | 81.2 (0.1) | 90.8 (0.1) | 96.8 (0.0) | 99.7 (0.0) | 98.0 (0.1) | 99.1 (0.0) | 99.7 (0.0) | 82.8 (0.1) | 93.5 (0.0) |
| | INT4 | 81.7 (0.6) | 91.5 (0.8) | 97.3 (0.5) | 99.8 (0.1) | 98.1 (0.2) | 99.2 (0.1) | 99.8 (0.1) | 83.5 (0.8) | 93.9 (0.4) |
| | INT3 | 82.2 (1.1) | 91.9 (1.2) | 97.3 (0.5) | 99.8 (0.1) | 98.1 (0.2) | 99.2 (0.1) | 99.7 (0.0) | 83.0 (0.3) | 93.9 (0.4) |
| | INT2 | 78.3 (-2.8) | 89.1 (-1.6) | 94.2 (-2.6) | 97.8 (-1.9) | 83.0 (-14.9) | 84.7 (-14.4) | 98.4 (-1.3) | 75.4 (-7.3) | 87.6 (-5.9) |
| | RTVQ | 79.1 (-2.0) | 89.0 (-1.7) | 94.5 (-2.3) | 97.9 (-1.8) | 91.9 (-6.0) | 93.8 (-5.3) | 99.3 (-0.4) | 76.8 (-5.9) | 90.3 (-3.2) |

Table F. Comprehensive task-level results for merging 8 classification tasks using ViT-L/14. For RTVQ, we use a 3-bit base vector and a 2-bit offset vector (equivalent to 2.375 bits per task).

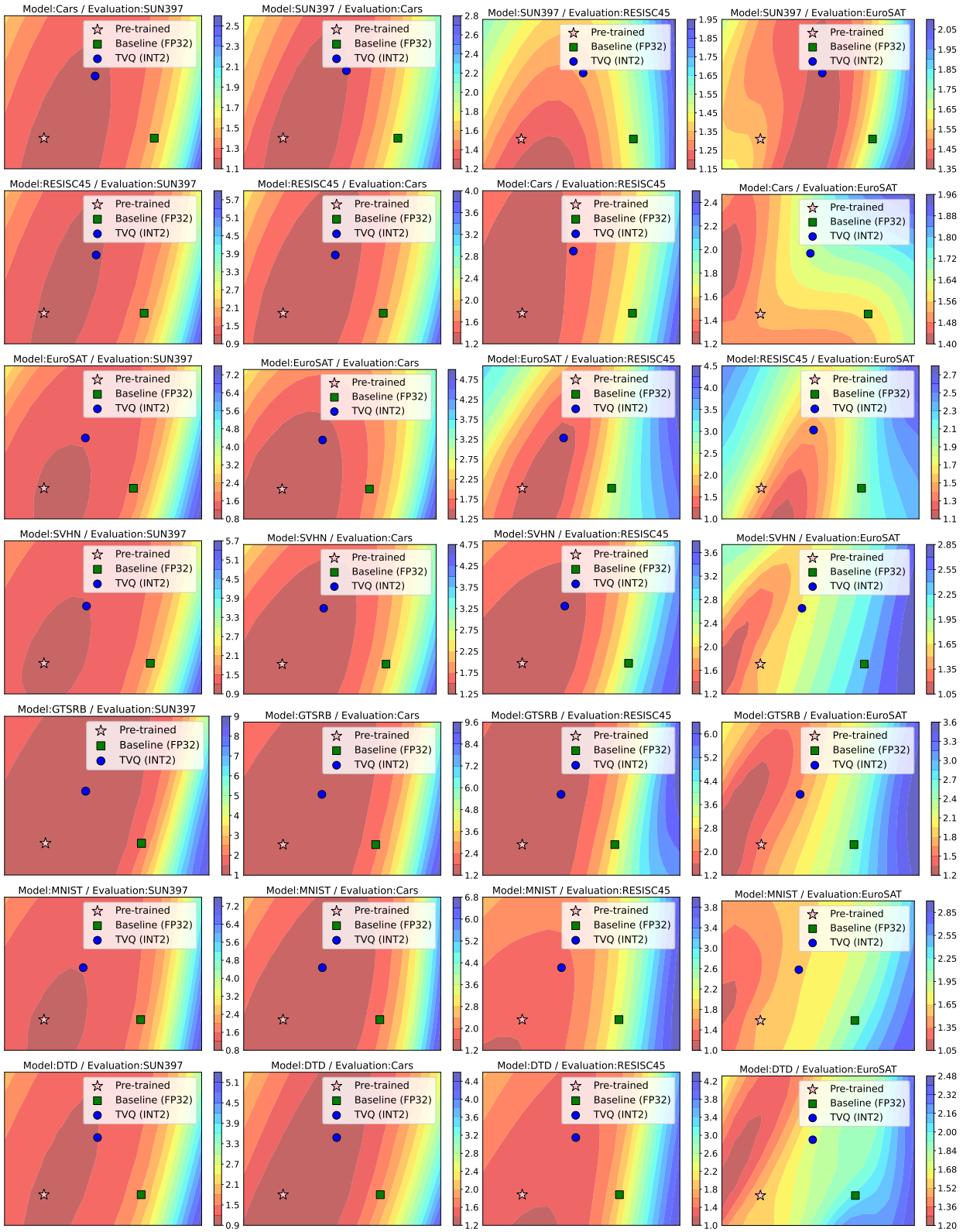



Figure F. Loss landscape visualization of cross task pairs for 2-bit TVQ. The results show evaluations on SUN397, Cars, RESISC45, and EuroSAT.

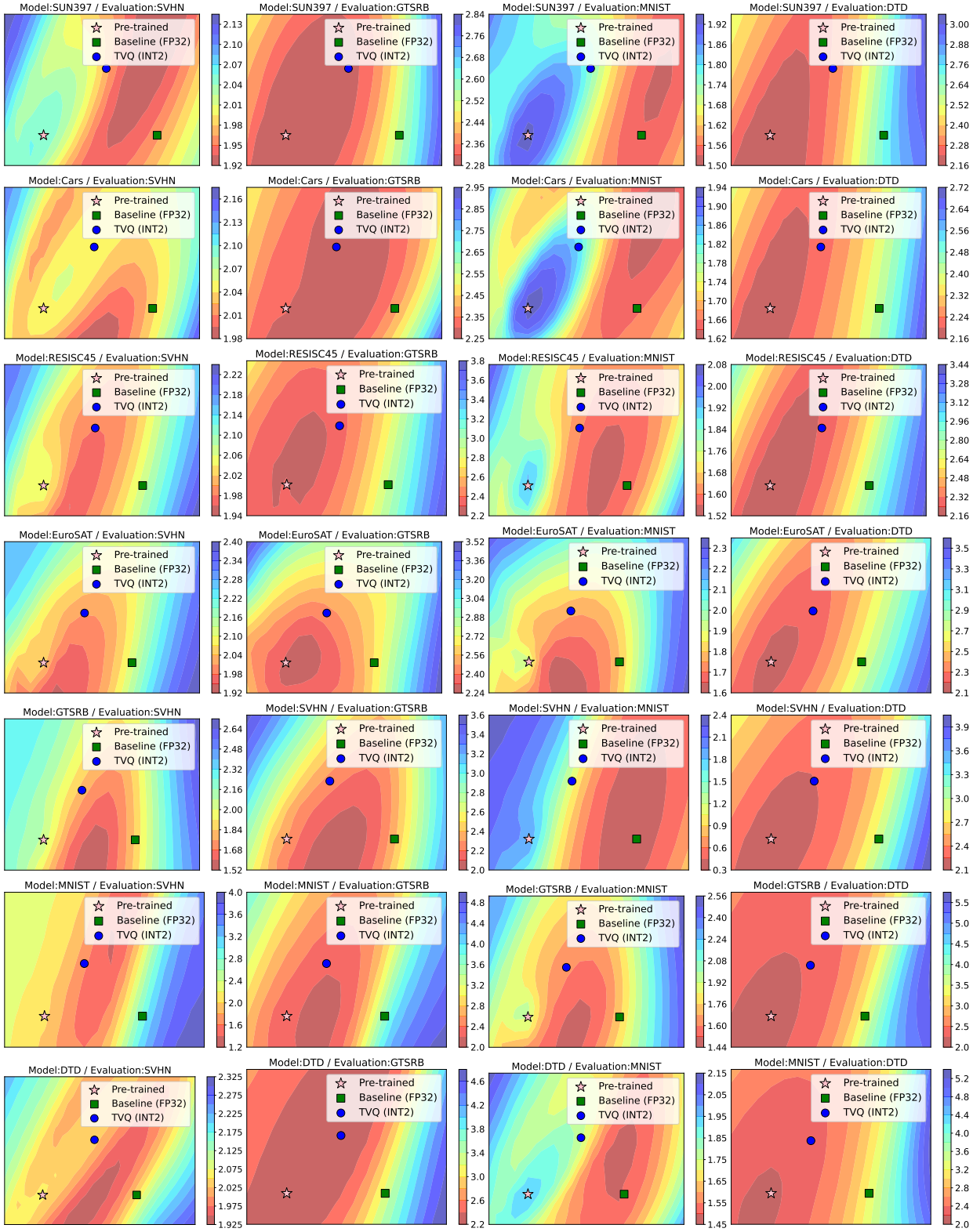



Figure G. Loss landscape visualization of cross task pairs for 2-bit TVQ. The results show evaluations on SVHN, GTSRB, MNIST, and DTD.

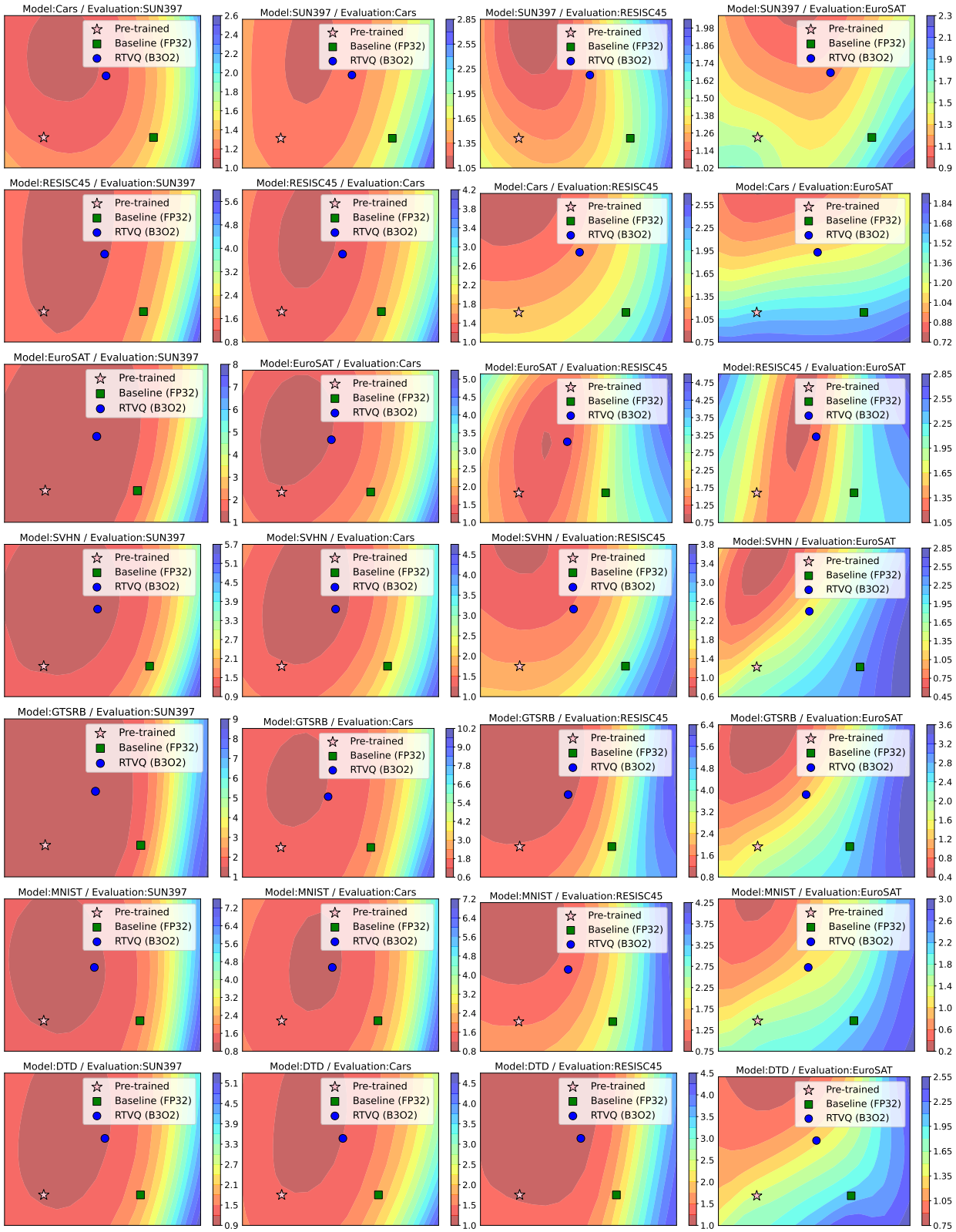



Figure H. Loss landscape visualization of cross task pairs for RTVQ (B2O3). The results show evaluations on SUN397, Cars, RESISC45, and EuroSAT.

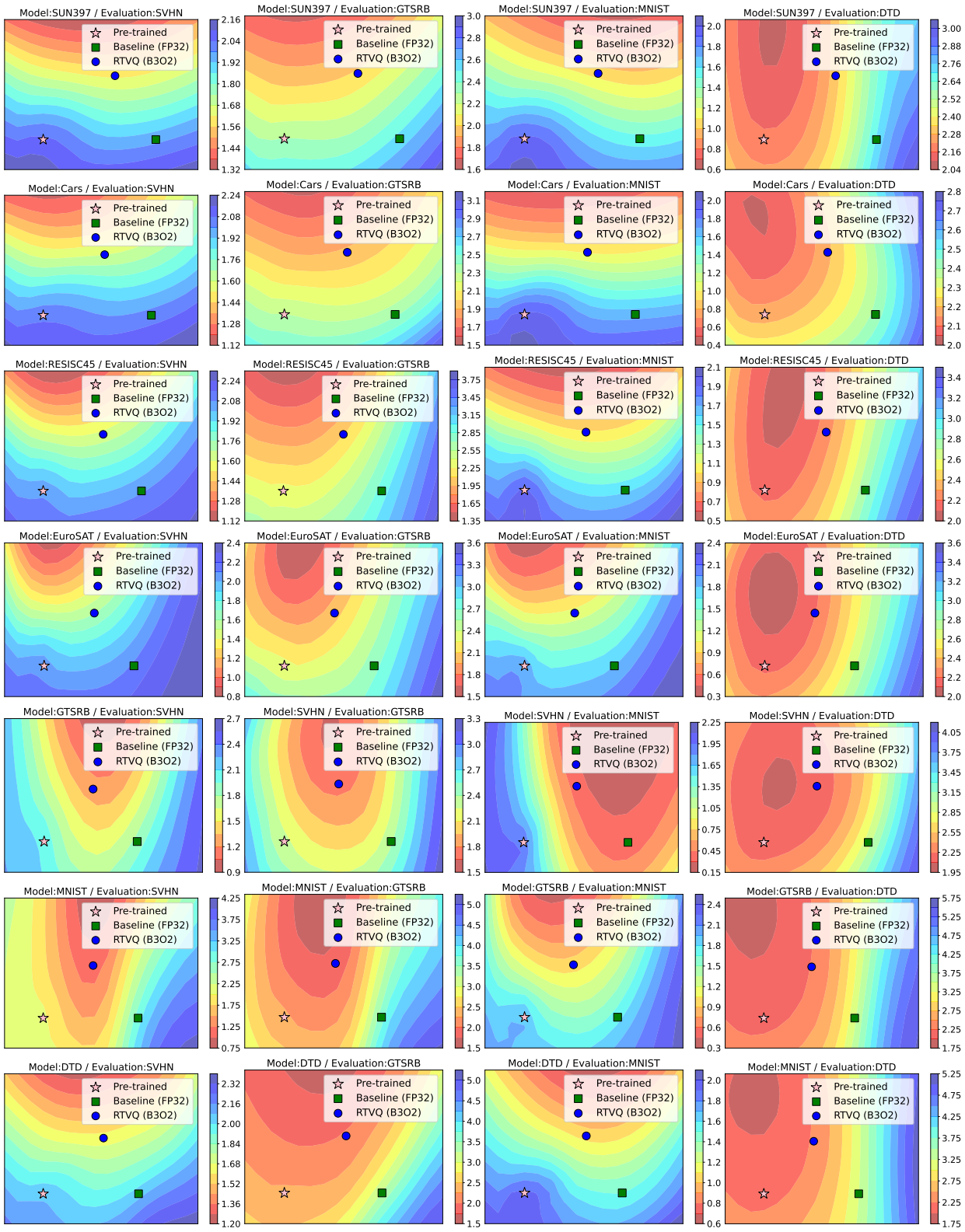



Figure I. Loss landscape visualization of cross task pairs for RTVQ (B2O3). The results show evaluations on SVHN, GTSRB, MNIST, and DTD.

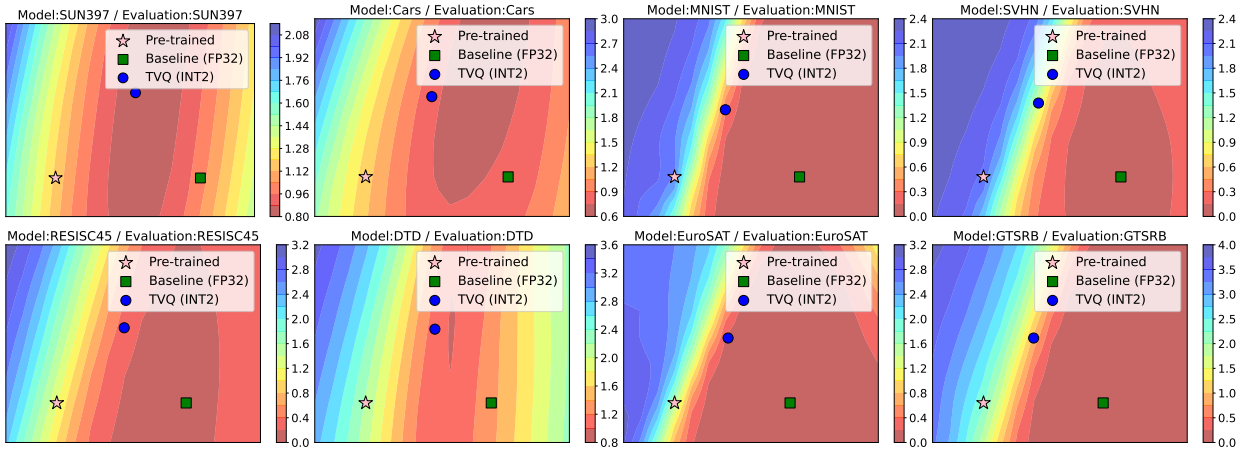



Figure J. Loss landscape visualization of all target task pairs for 2-bit TVQ.

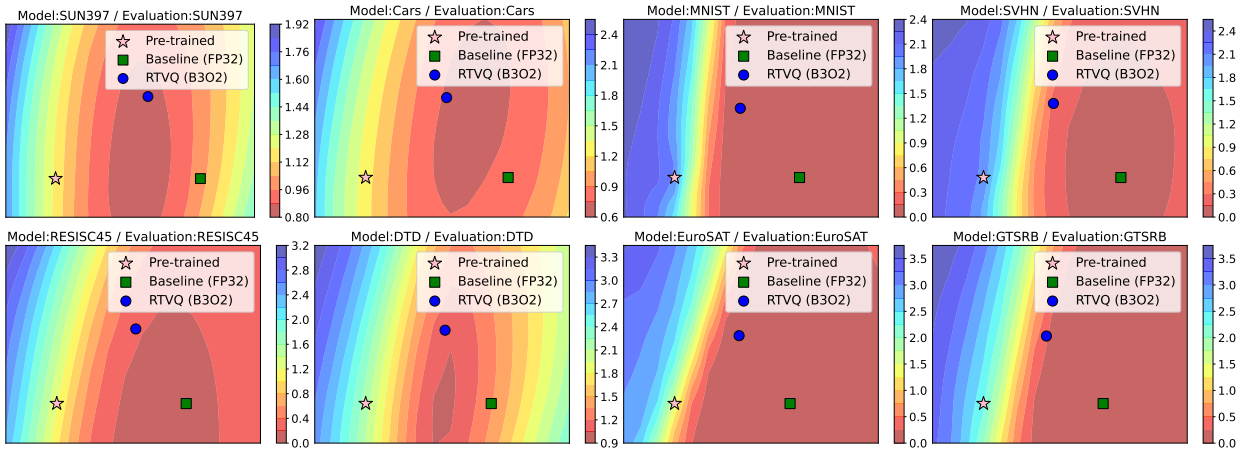



Figure K. Loss landscape visualization of all target task pairs for RTVQ (B2O3).

## References


[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 446–461. Springer, 2014. 1

[2] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. pages 1–14, Vancouver, Canada, 2017. Association for Computational Linguistics. 1

[3] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. 105(10):1865–1883, 2017. 1

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014. 1

[5] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018. 1

[6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proc. of Artificial Intelligence and Statistics (AISTATS)*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 1

[7] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *Proc. of International Joint Conference on Neural Networks*, pages 2921–2926. IEEE, 2017. 1

[8] MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 270–287. Springer, 2024. 2, 6

[9] Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proc. of Int'l Workshop on Paraphrasing (IWP)*, 2005. 1

[10] Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodola. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18695–18705, 2025. 2

[11] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Proc. of Neural Information Processing Systems (NeurIPS)*, 31, 2018. 3

[12] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. The third pascal recognizing textual entailment challenge. In *Proc. of ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, 2007. 1

[13] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Proc. of Int'l Conf. on Neural Information Processing (ICONIP)*, pages 117–124. Springer, 2013. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. 12(7): 2217–2226, 2019. 1

[16] Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging. *Proc. of Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2, 4, 6, 7, 8

[17] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *Proc. of Int'l Conf. on Learning Representation (ICLR)*, 2023. 1, 2, 4, 6, 7, 8

[18] Shankar Iyer, Nikhil Dandekar, Kornél Csernai, et al. First quora dataset release: Question pairs. data. quora. com. 2017. 1

[19] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proc. of Int'l Conf. on Computer Vision Workshops (ICCVW)*, pages 554–561, 2013. 1

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[21] Yann LeCun. The mnist database of handwritten digits. 1998. 1

[22] Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 1907. 1

[23] Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. Magmax: Leveraging model merging for seamless continual learning. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 379–395. Springer, 2024. 2, 6

[24] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *Proc. of Neural Information Processing Systems Workshops (NeurIPSW)*, page 4. Granada, 2011. 1

[25] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proc. of Indian Conf. on Computer Vision, Graphics and Image Processing (ICVGIP)*, pages 722–729. IEEE, 2008. 1

[26] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505. IEEE, 2012. 1

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 1

[29] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. 1

[30] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 746–760. Springer, 2012. 1

[31] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013. 1

[32] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *Proc. of International Joint Conference on Neural Networks*, pages 1453–1460. IEEE, 2011. 1

[33] Anke Tang, Li Shen, Yong Luo, Han Hu, Bo Du, and Dacheng Tao. Fusionbench: A comprehensive benchmark of deep model fusion. *arXiv preprint arXiv:2406.03280*, 2024. 1

[34] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Proc. of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 210–218. Springer, 2018. 1

[35] Alex Wang. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. of Int'l Conf. on Learning Representation (ICLR)*, 2019. 1

[36] Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jiménez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2024. 1, 2, 4, 7, 8

[37] Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, Francois Fleuret, and Pascal Frossard. Lines: Post-training layer scaling prevents forgetting and enhances model merging. In *Proc. of Int'l Conf. on Learning Representation (ICLR)*, 2025. 2, 4, 7, 8

[38] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics (TACL)*, 7:625–641, 2019. 1

[39] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. pages 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics. 1

[40] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 1

[41] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492. IEEE, 2010. 1

[42] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Proc. of Neural Information Processing Systems (NeurIPS)*, 36, 2024. 2, 4, 6, 7, 8

[43] Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *Proc. of Int'l Conf. on Learning Representation (ICLR)*, 2024. 2, 4, 7, 8