U-Net Transplant: The Role of Pre-training for Model Merging in 3D Medical Segmentation

Luca Lumetti*, Giacomo Capitani*, Elisa Ficarra, Simone Calderara, Costantino Grana, Angelo Porrello† ✉, and Federico Bolelli† ✉

University of Modena and Reggio Emilia, Italy
{*name.surname*}@unimore.it

Abstract. Despite their remarkable success in medical image segmentation, the life cycle of deep neural networks remains a challenge in clinical applications. These models must be regularly updated to integrate new medical data and customized to meet evolving diagnostic standards, regulatory requirements, commercial needs, and privacy constraints. Model merging offers a promising solution, as it allows working with multiple specialized networks that can be created and combined dynamically instead of relying on monolithic models. While extensively studied in standard 2D classification, the potential of model merging for 3D segmentation remains unexplored. This paper presents an efficient framework that allows effective model merging in the domain of 3D image segmentation. Our approach builds upon theoretical analysis and encourages wide minima during pre-training, which we demonstrate to facilitate subsequent model merging. Using U-Net 3D, we evaluate the method on distinct anatomical structures with the ToothFairy2 and BTCV Abdomen datasets. To support further research, we release the source code and all the model weights in a dedicated repository: `https://github.com/LucaLumetti/UNetTransplant`.

Keywords: Model Merging · Medical Segmentation · Task Vectors

1 Introduction

Although deep networks have achieved significant success in lesion segmentation and disease diagnosis [1,11], the segmentation of medical images still poses distinct challenges in obtaining high-quality annotated data. The scarcity of labeled data due to the time-intensive nature of manual annotations and the variability in imaging protocols across institutions makes it difficult to build robust models. As a result, fully annotated datasets are often unavailable at the outset of a project, and new diseases or segmentation classes may emerge later. In this respect, models deployed in real-world healthcare settings should ideally learn continuously while preserving previously acquired knowledge. A

*Equal contribution. Authors are allowed to list their name first on their CVs.
† Equal supervision. Authors are allowed to list their name last in their CVs.
✉ Corresponding authors: {angelo.porrello, federico.bolelli}@unimore.it.

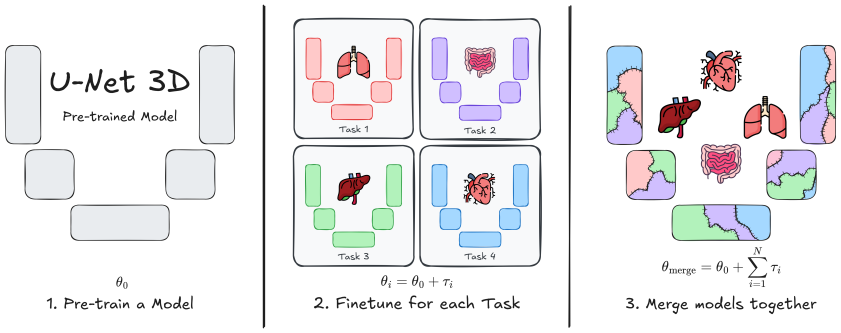

Fig. 1: Overview of model merging for 3D medical segmentation models.

straightforward approach for integrating new knowledge involves retraining the model from scratch on an aggregated dataset that includes both past and newly available data. However, strict privacy and security regulations may prohibit the long-term storage of patient records, and resources for full retraining may be unavailable, making this approach impractical or undesirable in medical contexts.

To support these scenarios, an ideal model should allow for fast and flexible adaptation, enabling the integration of new data or classes. If the model can accommodate novel anatomical structures without requiring re-training, it would reduce storage and deployment costs and potentially reduce the need for labeled data. From a medical perspective, removing the need for complete re-training would minimize the long-term storage of sensitive training data, simplify compliance with ethical committee requirements, and support a decentralized and modular development paradigm. Commercially, the ability to combine model capabilities without re-training would enable dynamic, client-specific software customization, thereby accelerating deployment and offering greater flexibility.

Notably, *model merging* permits updating and customizing AI models, facilitating knowledge transfer without full retraining [10,19,24,27]. Our approach builds on these foundations by utilizing task vectors [10,23], which represent modifications to a pre-trained model introduced during fine-tuning for a specific task. These vectors can be added to tune the model's functionalities (Fig. 1).

Unfortunately, model merging is not always practical, as it relies on the availability of effective pre-trained models. While standard computer vision tasks benefit from a wide selection of pre-trained base models, medical imaging—particularly tasks involving 3D segmentation—does not share the same advantage. In this respect, we aim to investigate the properties that a base pre-trained model must possess to ensure more effective model merging operations. Through both analytical and empirical assessments, we demonstrate why the base model should attain wide minima [28,29] in the optimization landscape. While wide minima have been investigated in continual learning [20,21] (*i.e.*, tasks succeed one after the other), their implications in the context of model merging—where models are integrated simultaneously—remain unexplored until this study.

Specifically, we present the first analysis of model merging for 3D image segmentation. Considering two well-known medical datasets (ToothFairy2 [2,3] and BTCV Abdomen [15]) and the standard 3D U-Net architecture, our study shows how models specialized for segmenting different anatomical structures can be successfully merged into a single model that can perform all the original tasks.

Contributions. We provide: *i)* an extensive analysis of model merging for 3D segmentation based on well-known medical datasets, revealing that combining task vectors is a flexible method for customizing models without re-training, *ii)* we offer both theoretical and empirical validation showing how a base model with a flat loss landscape enhances model merging, *iii)* alongside the source code, model's weights are publicly released to facilitate research.

2 Framework

We deal with a neural net $f(\cdot;\boldsymbol{\theta})$ designed for 3D segmentation, like 3D U-Net. The model has weights $\boldsymbol{\theta} \in \mathbb{R}^m$ and takes 3D images as input $\boldsymbol{x} \in \mathbb{R}^{H\times W\times D}$. The output is a 3D map of class distributions $p_\theta(\boldsymbol{y}|\boldsymbol{x})$, one for each voxel $\boldsymbol{y}$ in $\mathcal{Y} \in \mathbb{R}^{H\times W\times D\times C}$. We study a **multi-task learning framework** comprising T segmentation tasks, denoted as $\mathcal{T}$. Each task $t \in \mathcal{T}$ is associated with a dataset $\mathcal{D}_t$ of n_t training samples, sampled from a task-specific distribution $p_t(\boldsymbol{x}, \boldsymbol{y})$. Despite variations in these distributions (*e.g.*, different anatomical parts segmented in each task), all share a common loss function $\ell(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$ (*e.g.*, the cross-entropy loss), defined as the negative log-likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{x}, \mathcal{Y}) = -\sum_{\boldsymbol{y}\in\mathcal{Y}} \log p_\theta(\boldsymbol{y}|\boldsymbol{x})$.

Model Merging. To learn multiple segmentation tasks, we consider training a distinct set of weights for each task independently. We organize these models within a pool $\mathcal{P} = \{f(\cdot;\boldsymbol{\theta}_t) \mid \boldsymbol{\theta}_t \triangleq \boldsymbol{\theta}_0 + \boldsymbol{\tau}_t\}_{t\in\mathcal{T}}$ that can be expanded to accommodate for new tasks. Importantly, each model $f(\cdot;\boldsymbol{\theta}_t)$ is initialized from a shared set of **pre-trained weights** $\boldsymbol{\theta}_0$ and fine-tuned for its respective task. The displacement in weight space $\boldsymbol{\tau}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_0$ is called *task vector* [10] and, intuitively, it represents a direction in which the loss decreases for the t-th task.

As we discuss further, the models in the pool $\mathcal{P}$ can be selected and combined in arbitrary ways to construct a (personalized) multi-task model. The simplest approach to achieve this is by simply averaging the weights within the pool:

$$f_{\mathcal{P}} \triangleq f(\cdot;\boldsymbol{\theta}_{\mathcal{P}}) \quad \text{s.t.} \quad \boldsymbol{\theta}_{\mathcal{P}} = \boldsymbol{\theta}_0 + \textstyle\sum_{t=1}^{T} w_t \boldsymbol{\tau}_t, \quad \sum_{t=1}^{T} w_t = 1. \tag{1}$$

By adjusting the coefficients w_t, we can specialize the merged model toward specific tasks, deprioritizing others. Conversely, for a model that maintains a balance across all tasks, a uniform weighting scheme, $w_t = 1/T$, can be used.

The **goal** is to design an approach that learns and combines multiple 3D segmentation models, ensuring that the resulting merged model performs well across a combined set of tasks. To assess multi-tasking, we define the **empirical**

risk, *i.e.*, the average loss $\hat{\ell}(\boldsymbol{\theta}|\mathcal{D})$ over the union of all training tasks:[1]

$$\hat{\ell}(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{\sum_{t=1}^{T} n_t} \sum_{\boldsymbol{x},\boldsymbol{y} \in \bigcup_{t=1}^{T} \mathcal{D}_t} \ell(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}) \tag{2}$$

Research Question. While 2D image classification tasks can benefit from a variety of pre-trained models (*e.g.*, CLIP and DINO), 3D medical segmentation tasks face the absence of similar pre-trained models. In this respect, *how can we develop pre-trained models for 3D segmentation that facilitate model merging*?

2.1 Model Merging from a Pre-training Perspective

Following [22], we analyze model merging through the lens of the Taylor approximation of the loss function. Specifically, we indicate as $\ell_{\text{cur}}(\boldsymbol{\theta})$ the second-order approximation of the empirical risk, centered around the pre-trained weights $\boldsymbol{\theta}_0$:

$$\hat{\ell}_{\text{cur}}(\boldsymbol{\theta}) = \hat{\ell}(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla \hat{\ell}(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{3}$$

with $\nabla \hat{\ell}(\boldsymbol{\theta}_0) \triangleq \nabla_{\boldsymbol{\theta}} \hat{\ell}(\boldsymbol{\theta}_0)$ and $\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0) \triangleq \nabla^2_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_0)$ indicating the gradient and the Hessian around $\boldsymbol{\theta}_0$. Based on [22], assuming that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is a local minimum for the empirical risk $\hat{\ell}(\boldsymbol{\theta})$ across all tasks, the Hessian is positive semi-definite. It follows that the second-order approximation $\hat{\ell}_{\text{cur}}(\boldsymbol{\theta})$ of the empirical risk is locally convex. Utilizing Jensen's inequality (valid for convex functions) we can establish the following relationship between the merged model and the individuals:

$$\hat{\ell}_{\text{cur}}(\boldsymbol{\theta}_{\mathcal{P}} = \boldsymbol{\theta}_0 + \textstyle\sum_{t=1}^{T} w_t \boldsymbol{\tau}_t) \leq \sum_{t=1}^{T} w_t \; \hat{\ell}_{\text{cur}}(\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \boldsymbol{\tau}_t). \tag{4}$$

This inequality is informative because the term on the right offers a kind of worst-case upper limit for the performance of the merged model. In particular, the empirical risk $\hat{\ell}_{\text{cur}}(\boldsymbol{\theta}_{\mathcal{P}})$ of the merged model is constrained by the convex combination of the empirical risks associated with each individual model. This implies that if each individual model $\boldsymbol{\theta}_t$ performs accurately across all tasks, there are certain assurances regarding the risk level of the merged model $\boldsymbol{\theta}_{\mathcal{P}}$.

However, the issue with Eq. (4) is that, under a scenario with specialized models trained on separate tasks, we cannot ensure that each individual model $\boldsymbol{\theta}_t$ performs well across all tasks. Indeed, as $\boldsymbol{\theta}_t$ is trained exclusively on its specific distribution $p_t(\boldsymbol{x},\boldsymbol{y})$, its empirical risk is likely high for other data distributions $p_{t' \neq t}(\boldsymbol{x},\boldsymbol{y})$ ($\rightarrow$ **low out-of-distribution performance**). For this reason, the following augmented optimization problem was proposed [22] for the t-th learner:

$$\underset{\boldsymbol{\theta}_t}{\text{minimize}} \quad \mathbb{E}_{\boldsymbol{x},\boldsymbol{y} \sim p_t(\boldsymbol{x},\boldsymbol{y})} \left[\ell_{\text{cur}}(\boldsymbol{\theta}_t|\boldsymbol{x},\boldsymbol{y}) \right] + \mathcal{D}_{\text{KL}}(p_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}) || p_{\boldsymbol{\theta}_t}(\boldsymbol{y}|\boldsymbol{x})). \tag{5}$$

In essence, the out-of-distribution performance of each model is preserved through additional regularization provided by the term $\mathcal{D}_{\text{KL}}(\cdot)$, which acts explicitly on

[1] To simplify the notation, we will no longer explicitly denote the dependence of the loss on the data and write the individual loss and the empirical risk as $\ell(\boldsymbol{\theta})$ and $\hat{\ell}(\boldsymbol{\theta})$.

Table 1: Stable *vs.* plastic training regimes, metrics, and corresponding hyperparameters: Batch size (BS), Dropout (DO), and Learning rate (LR). λ_i correspond to the eigenvalues of $\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)$.

Regime	Dataset	BS	DO	LR	Dice↑	$\sum \boldsymbol{\lambda_i}\downarrow$	$\boldsymbol{\lambda_1}\downarrow$
Stable	Cui	4	0.5	10^{-3}	34.93	0.57	0.02
Plastic		8	0.0	10^{-4}	42.68	40.71	6.00
Stable	AMOS	4	0.5	10^{-3}	43.76	2.30	0.03
Plastic		8	0.0	10^{-4}	46.87	58.46	0.05

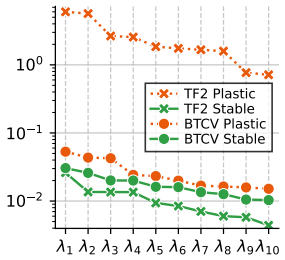

Fig. 2: Top 10 eigenvalues ↓.

out-of-distribution examples $\boldsymbol{x}, \boldsymbol{y} \sim p_{t' \neq t}(\boldsymbol{x}, \boldsymbol{y})$. The $\mathcal{D}_{\mathrm{KL}}(\cdot)$ term **aligns** the predictions $p_{\boldsymbol{\theta}_t}(\boldsymbol{y}|\boldsymbol{x})$ of the individual model $f(\cdot; \boldsymbol{\theta}_t)$ to those generated by the pre-trained model $\boldsymbol{\theta}_0$. By doing so, the individual model can achieve at least the performance level of the pre-trained model on external distributions $p_{t' \neq t}(\boldsymbol{x}, \boldsymbol{y})$, effectively reducing the upper bound on the right side of Eq. (4).

2.2 The Role of the Training Regime of the Pre-trained Model

While the authors of [22] drew inspiration from Eq. (5) to design a data-free regularization term, we take a different approach that avoids introducing explicit regularization. Instead, we focus on analyzing the roles of the training regime of the pre-trained model.

Thesis. We hypothesize that the tendency of the fine-tuned model $\boldsymbol{\theta}_t$ to retain pre-training knowledge is linked to the curvature of the pre-trained point $\boldsymbol{\theta}_0$ within the landscape of the empirical risk $\hat{\ell}(\cdot)$. To show that, we approximate the $\mathcal{D}_{\mathrm{KL}}(\cdot)$ as in [5]: if $\boldsymbol{\theta}_t - \boldsymbol{\theta}_0 \to 0$, the $\mathcal{D}_{\mathrm{KL}}(\cdot)$ term is close to the *distance* between $\boldsymbol{\theta}_t$ and the pre-training weights $\boldsymbol{\theta}_0$:

$$\mathcal{D}_{\mathrm{KL}}(p_{\boldsymbol{\theta}_0}(\boldsymbol{y}|\boldsymbol{x}) \;||\; p_{\boldsymbol{\theta}_t}(\boldsymbol{y}|\boldsymbol{x})) \approx \tfrac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0). \tag{6}$$

The weight distance is not isotropic but instead influenced by the Hessian of the empirical risk evaluated at $\boldsymbol{\theta}_0$. Thanks to Eq. (6) and the positive semi-definiteness of the Hessian around $\boldsymbol{\theta}_0$, we can establish a **bound** on $\mathcal{D}_{\mathrm{KL}}(\cdot)$:

$$\mathcal{D}_{\mathrm{KL}}(\dots) \approx \frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_t - \boldsymbol{\theta}_0) \leq \frac{1}{2}\lambda_1 \left\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\right\|^2 = \frac{1}{2}\lambda_1 \left\|\boldsymbol{\tau}_t\right\|^2, \tag{7}$$

where λ_1 is the **maximum eigenvalue** of the Hessian $\mathbf{H}_{\hat{\ell}}(\boldsymbol{\theta}_0)$ around the pre-training weights. The result is that the degradation in out-of-distribution performance relative to the pre-trained model is controlled by: *i)* the norm of the task vector, and *ii)* the maximum eigenvalue λ_1 of the Hessian. Notably, the entire spectrum of eigenvalues has been crucial in analyzing the geometry of the loss landscape and its impact on generalization capabilities [7,13]. Moreover, the maximum eigenvalue has been extensively used to characterize the width of a local minima [9,13,21]. In particular, a larger maximum eigenvalue suggests that the loss landscape is steeper along at least one dimension, which corresponds

Table 2: Details of the datasets used in our experiments. Data is not resampled, but it is preprocessed with z-score normalization and patch-based training.

Dataset	Modality	Volumes	Structs	Shape
AMOS [12] (**pre-training**)	CT	240	15	$148 \times 533 \times 560$
BTCV Abdomen [15]		30	13	$125 \times 512 \times 512$
Cui [6] (**pre-training**)	CBCT	151	42	$322 \times 402 \times 402$
ToothFairy2 [3]		480	42	$169 \times 356 \times 375$

to a *sharper minimum*. Conversely, smaller eigenvalues suggest *wider minima* because the surface of the loss function changes less drastically in those directions. Hence, to sum up, for a fixed task vector $\boldsymbol{\tau}_t$, the wider the curvature of the pre-trained model, the lower the loss in out-of-distribution performance during fine-tuning, and the better fine-tuned individual models will merge.

2.3 Biasing the Base Pre-Trained Model Towards Wide Minima

Building on this analytical finding, we propose modifying the training regime of the base pre-trained model to bias optimization toward wider minima. To do so, the approach is simple: inspired by [21], we act on some key hyperparameters—like batch size, dropout, and learning rate—that have been shown to affect generalization and the geometry of the minimum [8,16,25]. Following the terminology in [21], we define two distinct pre-training regimes, namely *stable* (wide minima) *vs. plastic* (sharp minima). The *stable* pre-training regime employs a small batch size, a higher learning rate, and increased dropout. In contrast, the *plastic* pre-training follows conventional self-supervised learning best practices, including as large as possible batch size, no dropout, and lower learning rates.

To analyze the effects of these hyperparameters, a preliminary result is reported in Tab. 1. We pre-train two base models (the one within the stable regime and the other in the plastic one) on two datasets for 3D medical image segmentation, namely AMOS [12] and Cui [6]. We then evaluate the average Dice on the corresponding test sets and compare the Hessian's eigenvalues as a proxy for the width of the pre-training optimum. Following [4], the Hessian's eigenspectrum is calculated with the trace of the empirical Fisher Information Matrix (FIM) [14], as a (diagonal) approximation of the intractable Hessian. As observed, the performance of the two base models (stable *vs.* plastic) is comparable across both datasets; however, the stable model achieves a remarkably lower trace (Fig. 2). This indicates that manipulating hyperparameters is a simple yet effective way to influence the geometry of the solution attained by the pre-trained model.

3 Experiments and Results

Datasets and Task Splits. Considering four public datasets, we categorize experiments into two settings based on the target anatomical regions: *i) abdominal datasets* (AMOS [12] and BTCV Abdomen [15]) and *ii) maxillofacial*

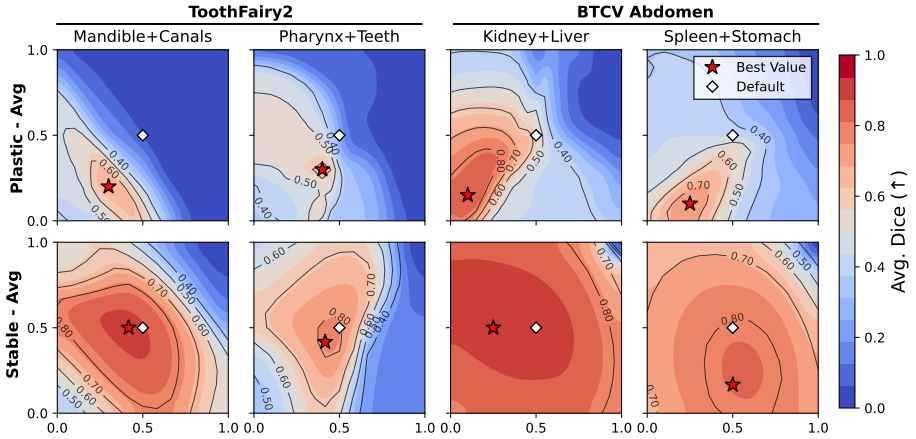

Fig. 3: The average Dice score for two classes when merging task vectors τ_1 (x-axis) and τ_2 (y-axis) by varying w_1 and w_2. The star (★) marks the maximum Dice score, while the diamond (◇) denotes the default w_i values. The first row shows task vectors from *plastic* pre-training and the second row from *flat* pre-training, both merged using average. All the plots have the same x and y scales.

datasets (Cui [6] and ToothFairy2 [3,18]). The summary characteristics are provided in Tab. 2. In the abdominal scenario, we use AMOS for pre-training and four BTCV classes (Liver, Spleen, Kidney, and Stomach) to create four tasks. In the maxillofacial scenario, we use Cui for pre-training and ToothFairy2 for fine-tuning, with four tasks based on Mandible, Pharynx, Teeth, and Canals.

Training. We perform stable and plastic pre-training for both AMOS and Cui according to the setup in Tab. 1. To perform fine-tuning, we replace the final $1 \times 1 \times 1$ convolution with a new one; for the rest of the layers, we fine-tune the corresponding parameters $\boldsymbol{\theta}_0$ through a task vector $\boldsymbol{\tau}_t$ (initialized at zero). We optimize with *AdamW* [17] and a weight decay penalty of 0.1 to discourage large task vector norms. Training runs for 10 epochs.

3.1 Impact of Pre-Training Regime on Model Merging

In each plot of Fig. 3, we consider a pair of tasks (*e.g.*, Mandible + Canals) and evaluate the Dice score of the merged model while varying merging coefficients w_1 and w_2. By comparing plastic (first row) *vs.* stable (second row) pre-training, we can say that stable pre-training allows for remarkably robust performance, exhibiting lower sensitivity to the merging coefficients—a feature that, in real-world applications, reduces the overhead associated with hyperparameter tuning. As further proof, for the stable regime the uniform weighing scheme ◇ ($w_{1,2} = 0.5$) is always closer to the best configuration ★ (found by hyperparameter tuning on an held-out set).

After examining a scenario where pairs of tasks are merged, we extend our analysis to a setting with four task vectors. We report in Fig. 4 the results (Dice

Table 3: Performance scores obtained from pairwise task vector merging.

Dataset	w	Merging Strategy	Spl. Kid.	Spl. Liv.	Spl. Sto.	Kid. Liv.	Kid. Sto.	Liv. Sto.	Avg.
BTCV Abdomen	Default ◇	Average [10]	91.41	92.18	80.61	90.85	77.18	80.91	85.52
		TIES [26]	82.80	90.76	76.83	88.69	58.56	76.69	79.05
	Best ★	Average [10]	92.64	92.22	82.09	91.01	78.97	81.80	86.45
		TIES [26]	92.42	91.88	81.55	91.07	77.72	81.04	85.95
	-	Joint	91.40	93.34	78.38	92.31	91.79	88.86	89.35

Dataset	w	Merging Strategy	Mand. IACs	Mand. Teeth	Mand. Phar.	IACs Teeth	IACs Phar.	Teeth Phar.	Avg.
ToothFairy2	Default ◇	Average [10]	89.54	82.55	87.70	56.08	57.08	81.27	75.70
		TIES [26]	88.70	79.89	88.96	58.51	63.44	73.72	75.54
	Best ★	Average [10]	89.67	82.78	91.89	68.16	75.19	81.27	81.49
		TIES [26]	89.24	82.33	91.97	67.94	68.91	81.07	80.24
	-	Joint	98.75	97.33	98.26	83.04	97.10	93.61	94.68

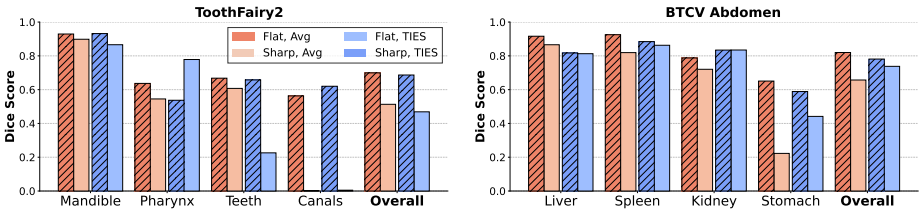

Fig. 4: Task-wise performance after merging four distinct task vectors with weight averaging and TIES. The Overall bars aggregate results across tasks.

score) on each task separately and also the average (**Overall**). Beyond comparing stable ◪ *vs.* plastic ◪ pre-training, we also examine their impact on TIES Merging [26], a well-established alternative to uniform averaging. The results in Fig. 4 show that, in both settings, the performance of the merged model is primarily influenced by the type of pre-training rather than the merging method. This is evidenced by the significant performance gains achieved with stable pre-training (*e.g.*, with uniform averaging, yielding an improvement of +18.60 on ToothFairy2 and +16.28 on BTCV).

Further Comparative Analysis. To assess the effectiveness of model merging for 3D segmentation, we include a reference approach representing re-training from scratch, where the pre-trained model is fine-tuned on both classes jointly. As shown in Tab. 3, in BTCV Abdomen, Kidney+Stomach shows the largest drop w.r.t. the joint training ($\sim$ 18.60 Dice score), while other pairs achieve similar performance, indicating effective merging. In contrast, the gap is significantly larger in ToothFairy2, likely due to greater variation in the shape, size, and

intensity values of maxillofacial structures. We conjecture that such an increased variability leads to higher interference when merging the relative task vectors.

4 Conclusion

We pioneer model merging for 3D segmentation, showing the pivotal role of the training regime of the base U-Net model. Based on our findings, the life cycle of many existing models could be revised in favor of modular paradigms. Future work will investigate scalability to larger task sets and novel merging strategies.

Acknowledgments. This work was supported by the Italian Ministerial grants PRIN 2022: "B-Fair: Bias-Free Artificial Intelligence Methods for Automated Visual Recognition" (CUP E53D23008010006) and by the University of Modena and Reggio Emilia and Fondazione di Modena through the "Fondo di Ateneo per la Ricerca - FAR 2024" (CUP E93C24002080007) and FARD-2024. The work also received funding from DECIDER, the European Union's Horizon 2020 research and innovation programme under GA No. 965193 and "AIDA: explAinable multImodal Deep learning for personAlized oncology" (Project Code 20228MZFAA). We acknowledge the CINECA award under the ISCRA initiative for providing high-performance computing resources.

Disclosure of Interests. The authors have no conflicts of interest to declare.

References

1. Aggarwal, R., Sounderajah, V., Martin, G., Ting, D.S., Karthikesalingam, A., King, D., Ashrafian, H., Darzi, A.: Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digital Medicine **4**(1) (2021)
2. Bolelli, F., Lumetti, L., Vinayahalingam, S., Di Bartolomeo, M., Pellacani, A., Marchesini, K., Van Nistelrooij, N., Van Lierop, P., Xi, T., Liu, Y., et al.: Segmenting the Inferior Alveolar Canal in CBCTs Volumes: the ToothFairy Challenge. IEEE Transactions on Medical Imaging (2024)
3. Bolelli, F., Marchesini, K., van Nistelrooij, N., Lumetti, L., Pipoli, V., Ficarra, E., Vinayahalingam, S., Grana, C.: Segmenting Maxillofacial Structures in CBCT Volumes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
4. Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., Zecchina, R.: Entropy-SGD: Biasing Gradient Descent into Wide Valleys. Journal of Statistical Mechanics: Theory and Experiment **2019**(12) (2019)
5. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence. In: Proceedings of the European Conference on Computer Vision (2018)
6. Cui, Z., Fang, Y., Mei, L., Zhang, B., Yu, B., Liu, J., Jiang, C., Sun, Y., Ma, L., Huang, J., et al.: A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. Nature Communications **13**(1) (2022)
7. Dinh, L., Pascanu, R., Bengio, S., Bengio, Y.: Sharp Minima Can Generalize For Deep Nets. In: International Conference on Machine Learning (2017)

8. Frankle, J., Schwab, D.J., Morcos, A.S.: The Early Phase of Neural Network Training. In: International Conference on Learning Representations (2020)
9. Hochreiter, S., Schmidhuber, J.: Flat Minima. Neural Computation **9**(1) (1997)
10. Ilharco, G., Ribeiro, M.T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., Farhadi, A.: Editing models with task arithmetic. In: International Conference on Learning Representations (2023)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2) (2021)
12. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: AMOS: ALarge-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. Neural Information Processing Systems **35** (2022)
13. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In: International Conference on Learning Representations (2017)
14. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13) (2017)
15. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5 (2015)
16. Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., Gur-Ari, G.: The large learning rate phase of deep learning: the catapult mechanism. arXiv preprint arXiv:2003.02218 (2020)
17. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: International Conference on Learning Representations (2019)
18. Lumetti, L., Pipoli, V., Bolelli, F., Ficarra, E., Grana, C.: Enhancing Patch-Based Learning for the Segmentation of the Mandibular Canal. IEEE Access (2024)
19. Matena, M.S., Raffel, C.A.: Merging Models with Fisher-Weighted Averaging. Advances in Neural Information Processing Systems (2022)
20. Mehta, S.V., Patil, D., Chandar, S., Strubell, E.: An Empirical Investigation of the Role of Pre-training in Lifelong Learning. Journal of Machine Learning Research **24**(214) (2023)
21. Mirzadeh, S.I., Farajtabar, M., Pascanu, R., Ghasemzadeh, H.: Understanding the Role of Training Regimes in Continual Learning. Advances in Neural Information Processing Systems (2020)
22. Porrello, A., Bonicelli, L., Buzzega, P., Millunzi, M., Calderara, S., Cucchiara, R.: A Second-Order Perspective on Model Compositionality and Incremental Learning. In: International Conference on Learning Representations (2025)
23. Rinaldi, F., Capitani, G., Bonicelli, L., Crisostomi, D., Bolelli, F., Ficarra, E., Rodolà, E., Calderara, S., Porrello, A.: Update Your Transformer to the Latest Release: Re-Basin of Task Vectors. In: International Conference on Machine Learning (2025)
24. Tam, D., Bansal, M., Raffel, C.: Merging by Matching Models in Task Parameter Subspaces. Transactions on Machine Learning Research (2024)
25. Xie, Z., Sato, I., Sugiyama, M.: A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast. arXiv preprint arXiv:2002.03495 (2020)

26. Yadav, P., Tam, D., Choshen, L., Raffel, C.A., Bansal, M.: TIES-Merging: Resolving Interference When Merging Models. In: Neural Information Processing Systems (2024)
27. Yang, E., Wang, Z., Shen, L., Liu, S., Guo, G., Wang, X., Tao, D.: AdaMerging: Adaptive Model Merging for Multi-Task Learning. In: International Conference on Learning Representations (2024)
28. Yao, Z., Gholami, A., Lei, Q., Keutzer, K., Mahoney, M.W.: Hessian-based Analysis of Large Batch Training and Robustness to Adversaries. In: Neural Information Processing Systems (2018)
29. Yu, F., Qin, Z., Liu, C., Zhao, L., Wang, Y., Chen, X.: Interpreting and Evaluating Neural Network Robustness. In: International Joint Conference on Artificial Intelligence (2019)