InterConv: Set Interaction for Improved Biomedical Image Segmentation

Anonymous CVPR submission

Paper ID 13795

Abstract

Image segmentation is an important part of many biomedical research and clinical pipelines. Because images within a dataset are often similar in appearance and composition, structures in one image can contain information that is useful for segmenting other images. However, existing image segmentation models segment each input image independently, limiting their ability to share this information.

We present InterConv, a mechanism that enables segmentation models to interact and share information across a set of structurally related images. InterConv is a layer that can be inserted within any network to facilitate set interaction with intermediate sample features without changing the fundamental network architecture, and therefore can be integrated into most existing segmentation models. We demonstrate the effectiveness of InterConv by applying it to two state-of-the-art image segmentation architectures: UNets and Vision Transformers, and tackle challenging tasks in both automatic and interactive biomedical image segmentation. By learning to interact samples through aggregated set features, InterConv consistently improves per-sample segmentation performance, sometimes by up to 19%.

1. Introduction

Image segmentation is a core step in many biomedical image analysis pipelines. Machine learning-based models have been shown to be useful for segmenting images of a wide range of modalities, spanning across diverse biomedical domains, Typically, these models segment one image at a time. However, in many situations there are multiple images acquired using the same modality and of the same anatomy, resulting in similar structural compositions among samples. In some cases, the images have only minor differences, for example, longitudinal scans of the same subject. In the same way that pixels can provide contextual information about neighboring pixels, samples from a set of related images can potentially provide useful information about each other.

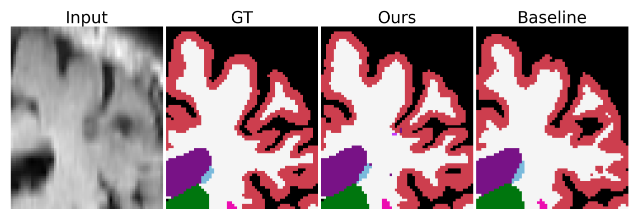


Figure 1. **InterConv helps models achieve substantially higher segmentation accuracy, as seen on the cortex (red)**. From **left** to **right** we show a close-up view of an input T2-FLAIR scan, with corresponding ground truth segmentation map, prediction with our proposed model, and the baseline prediction.

We present a novel learning-based segmentation mechanism, InterConv, that enables a model to share information across the set by interacting a set of related input images and making prediction for each of the sample. Our contributions are:

- We introduce InterConv, a novel layer that is easily incorporated in most modern network designs to enable interaction among input images.
- We demonstrate the generalizability of InterConv by applying it to two different state-of-the-art segmentation architecture paradigms: convolutional UNets and Vision Transformers.
- We evaluate the effectiveness of InterConv in two segmentation scenarios: (1) *interactive* segmentation of diverse biomedical image datasets, (2) *automatic* segmentation of low-resolution MRI brain scans. We find that InterConv improves segmentation accuracy in both scenarios.

2. Related Work

Medical Image Segmentation. Image segmentation is widely-studied in many biomedical domains. Most of today's existing methods use a convolutional neural network

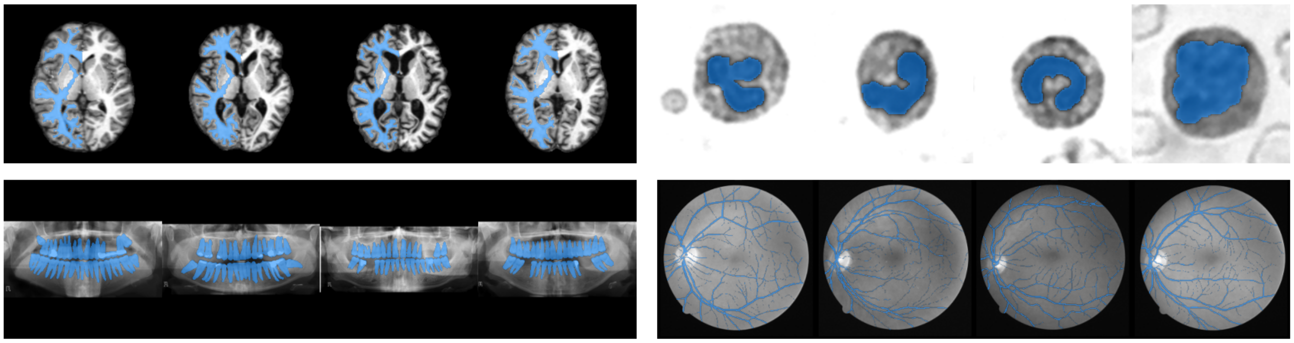


Figure 2. **Examples of biomedical image segmentation tasks with spatial consistency**. The ground truth segmentation is overlayed in blue [1, 45, 81, 107, 120].

to produce label maps for a single image [9, 23, 29, 48, 51, 54, 98]. When well-designed, UNet-like convolutional architectures [52, 73, 105] continue to perform similarly to more recent and more complex architectures, based on transformers [16, 37, 114] or state-space [79, 112] models. All of these methods make predictions for a single input image, independent of other images in the test set. In contrast, our work provides a mechanism to interact intermediate sample features in all these networks, and improve segmentation accuracy on each of the samples.

Multimodal Image Inputs. Some methods integrate information from multiple modalities to improve biomedical segmentation results [40, 57, 86]. These networks use a fixed number of input channels [17, 27, 28], which restricts the amount of input information the network can integrate [18, 20, 50, 51, 54, 91, 118]. Importantly, the multimodal inputs are related to a single sample and the networks make independent predictions on each sample. We propose a mechanism that is agnostic to the number of inputs in the set and produces a separate prediction for each set entry.

Multiple Predictions. Some segmentation methods generate multiple predictions for a given image and aggregate them, to improve accuracy or estimate uncertainty [51, 96, 103, 108, 110]. Some stochasatic segmentation methods are designed to model the variability in multiple raters [6, 61, 95, 96, 115]. While these methods produce, and sometimes interact among, multiple segmentations, they all operate on a single input image.

Data-driven (non-parametric) Mechanisms. Data-driven (or non-parametric) methods yield improved performance with higher amount of data available at inference. Recent deep-learning methods have combined parametric neural networks with data-driven mechanisms, to draw on data available at inference. For example, *in-context learning* methods employ a neural network that adapts to new tasks based on a *context*, often provided as a flexible set of example input-output pairs [12, 21, 96, 111]. In InterConv we also propose a non-parametric mechanism for jointly segmenting a set of multiple samples, and to enable interaction across the set. Moreover, InterConv can handle input sets of variables sizes.

Interactive Segmentation. Medical researchers and clinicians often need to perform *new* segmentation tasks involving new images, modalities, or labels that require them to perform manual segmentation. Interactive segmentation models reduce this burden by using sparse annotations, such as clicks and scribbles, as an additional model input. Recent foundation models for interactive segmentation yield promising results on many medical image segmentation tasks, but can perform poorly on tasks unseen during training, due to ambiguity in desired segmentation targets because of sparse annotations [60, 78, 113].

We demonstrate that with InterConv such ambiguities can be more easily resolved through set interaction where annotations in one image can be transferred across the set and help to segment other images. We find that this setting improves segmentation accuracy and requires substantially less total user effort than existing methods.

3. InterConv

Standard deep learning approaches to segmentation learn a function $y = g_\theta(x)$, with parameters θ, that outputs a prediction y given a single input image x.

We propose InterConv, a mechanism that, when applied to an existing network, enables it to model a function $\mathcal{Y} = f_\theta(\mathcal{X})$ that jointly predicts a set of outputs $\mathcal{Y} = \{y_i\}$ given a set $\mathcal{X} = \{x_i\}$ of input samples x_i, where each element y_i in the output set $\mathcal{Y}$ is the prediction corresponding to input image x_i. The input set $\mathcal{X}$ can be of flexible size. Input samples interact through the *InterConv Layer*, so that information from each sample can be shared and contribute to all predictions of the input set.

InterConv Layer. The InterConv Layer enables sample interaction across the input set by taking in a set of individual

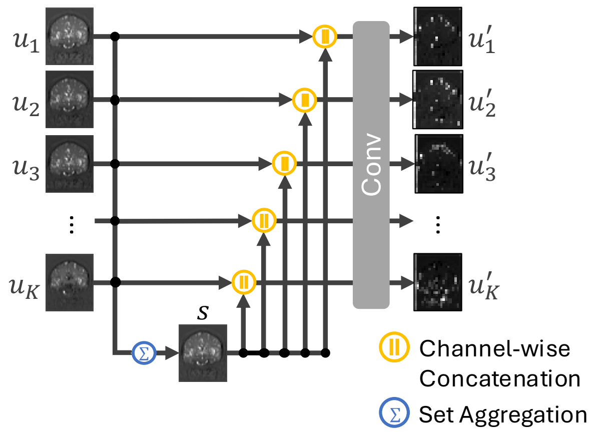


Figure 3. **Illustration of InterConv Layer.** InterConv Layer takes a set of intermediate features $\{u_i\}_{i=1}^{K}$ as input, aggregates them to produce the set context $\mathbf{s}$, and fuses the set context representation with original sample features by concatenation and applying a feed-forward layer (FFW).

sample features $\mathcal{U}$ from a given network layer, and producing an updated set of features for each sample $\mathcal{U}'$ after set interaction. The InterConv Layer first summarizes interacted information in a set context feature $\mathbf{s}$ through the *set aggregation* step, and combines it with individual information through the *feature fusion* step.

Set Aggregation. InterConv first takes a set of individual image representations $\mathcal{U} = \{\mathbf{u_i}\}$, where each $\mathbf{u_i}$ is an intermediate feature representation that corresponds to input image x_i and has size $F \times h \times w$, with height h and width w and F features. The set aggregation step computes set context $\mathbf{s}$ as a pixel-wise average across samples for each feature,

$$\mathbf{s} = \frac{1}{K}\sum_{i=1}^{K} \mathbf{u_i}. \tag{1}$$

Feature Fusion. Given the set context $\mathbf{s}$ and image-specific feature $\mathbf{u_i}$, InterConv first applies a channel-wise concatenation:

$$\mathbf{c_i} = \texttt{Concat}(\mathbf{u_i}; \mathbf{s}), \tag{2}$$

then applies a shared convolution layer to each sample feature $\mathbf{c_i}$ in the set to integrate the interacted information with the individual sample features

$$\mathbf{u'_i} = \texttt{Conv}(\mathbf{c_i}). \tag{3}$$

The final output features $\mathbf{u'_i}$ are the same size as the output from the original network layer, enabling InterConv to be used after any existing intermediate layer (Fig. 3).

Training. With InterConv, the network makes prediction on the entire set of images $\mathcal{X}$, and its weights are optimized jointly on the whole set:

$$\mathcal{L}_\theta(\mathcal{D}) = \mathbb{E}_K\left[\mathbb{E}_{\mathcal{X},\mathcal{Y}\subset\mathcal{D}}\left[\mathcal{L}\left(\hat{\mathcal{Y}}, \mathcal{Y}\right)\right]\right] \tag{4}$$

$$= \mathbb{E}_K\left[\mathbb{E}_{\mathcal{X},\mathcal{Y}\subset\mathcal{D}}\left[\sum_{i=1}^{K}\mathcal{L}_{seg}\left(\hat{y}_i, y_i\right)\right]\right] \tag{5}$$

where $\mathcal{D}$ is a dataset containing sets of image samples $\mathcal{X}$ and targets $\mathcal{Y}$ with various set sizes K, and $\hat{\mathcal{Y}}$ is the collection of network predictions on $\mathcal{X}$. $\mathcal{L}$ is the desired segmentation loss over the input set, which is the sum of per-sample segmentation loss $\mathcal{L}_{seg}$.

Inference. Once trained, any network with the InterConv layer added can perform joint prediction on input image sets of any size.

4. Experiments

We assess the ability of InterConv to productively use information from multiple input images. To do so, we explore two common applications where it is challenging to construct accurate segmentation maps from single input images: 1) interactive segmentation and 2) automatic segmentation of the low-quality clinical-grade MRI scans that are prevalent in clinical practice.

Baselines. We demonstrate the effectiveness of InterConv integrating it into two of the most widely used image segmentation frameworks: convolutional UNets and Vision Transformers. Specifically, we apply InterConv to the following architectural designs:

- SimpleUNet, a UNet-like architecture for general image segmentation with a single convolution block at each encoder and decoder step,
- nnUNet [51], a UNet-based biomedical image segmentation network that automatically chooses some network aspects depending on the dataset [51, 52],
- SwinUNet [14], a transformer-based state-of-the-art medical image segmentation model, and
- ScribblePrompt-UNet [113], a UNet-based state-of-the-art model for interactive segmentation.

Details about how we integrate InterConv into these architectures appear below.

Evaluation. We evaluate model predictions using Dice Score [26] averaged across foreground classes.

4.1. Interactive Segmentation

We first evaluate InterConv in the setting of interactive segmentation, focusing on segmentation tasks not seen during training.

Setup. Given an input image x_i and a collection of interactions z_i, we predict binary segmentation y_i. With InterConv layers, the model *jointly* segments a set of K images of sim-

CVPR
#13795

ilar anatomy $\{x_i\}_{k=1}^{K}$, given simulated prompts $\{z_i\}_{k=1}^{K}$, and produces a set of segmentation $\{\hat{y}_i\}_{k=1}^{K}$ for each input.

Interactions. Following recent methods [113], we consider interactive segmentation using scribbles and clicks. We simulate positive prompts from label area y_i and negative prompts from $1 - y_i$, following the generation method of ScribblePrompt:

- Line scribbles: We randomly sample end points from the label and background areas, create a line segment that joins the end points, and warps the line segment to make sure that all points on the deformed path falls into one of the label or background categories.
- Random clicks: We randomly sample pixels that belong to the label or background area.

Network Architectures. We follow the architecture of ScribblePrompt-UNet [113], a state-of-the-art interactive segmentation model with an efficient UNet-based architecture with 192 features at each convolution layer. We construct ScribblePrompt-InterConv by inserting a InterConv Layer with kernel size 3 at the end of each convolution layer of the baseline network. We proportionally downscale feature sizes in ScribblePrompt-InterConv to 137, so that both networks possess a similar number of parameters ($\sim$3.6M).

Data. We use the datasets in MegaMedical [12, 96, 113]: 74 diverse biomedical image datasets, including over 48,000 images, 16 image types, and 602 labels.

We partition the datasets into 65 training datasets and 9 evaluation datasets using the same splits as [113]. Each evaluation dataset is partitioned into validation data, used for model selection, and test data, used for final evaluation. We report results on the test splits of the 9 evaluation datasets, unseen by the models during training [1, 2, 7, 36, 67, 94, 107, 119, 120].

We define a task as a combination of dataset, modality, axis, and label. For multi-label datasets we consider each label as a separate binary segmentation task. For 3D datasets, we use the middle slice and slice with maximum label area to create 2D tasks.

Set Grouping. We sample sets containing different images from the same segmentation task. This aligns with a realistic scenario, as users often need to segment several images from the same dataset.

Training. *Training Objective*. We minimize Eq. 5 where $\mathcal{L}_{seg}$ is a combination of Soft Dice Loss [87] and Focal Loss [71] using the Adam optimizer [58].

Prompt generation. We train all models with 2 prompt types: scribbles and clicks. For each prompt type we sample a random number of positive prompts ranging from 1 to 6, and a random number of negative prompts ranging from 0 to 6.

We train all models with fixed batch size of 8 sets, with set sizes $K \sim Cat[1, 8]$ at each iteration. Following the training procedure in ScribblePrompt, we train both ScribblePrompt-UNet and ScribblePrompt-InterConv for 20,000 epochs with 125 batches per each epoch.

Evaluation. For evaluation, we use the same prompt types used during training. We evaluate model performance using Dice Score [26] on the 9 held-out datasets. We report results averaged across 5 runs with different (random) prompt initializations. At inference, we vary set size $K \in \{1, 2, 4, 8\}$ and the number of prompt interactions $n_{pos}, n_{neg} \in \{1, 2, 4, 8\}$. We filter out tasks with fewer than 8 examples to make sure that scan images are distinct within a set, and that evaluations are performed on the same collection of test scans for each set size.

Results. Fig. 4 shows that for both scribbles and clicks, with all prompt sizes, ScibblePrompt- outperforms the baseline for set sizes greater than 1. These improvements are significant with a pairwise t-test (p-value ¡ 7e-4). The improvement in Dice grows larger with larger set sizes. Fig. 5 provides illustrative examples highlighting this improvement. More interactive segmentation visualizations can be found in the Appendix.

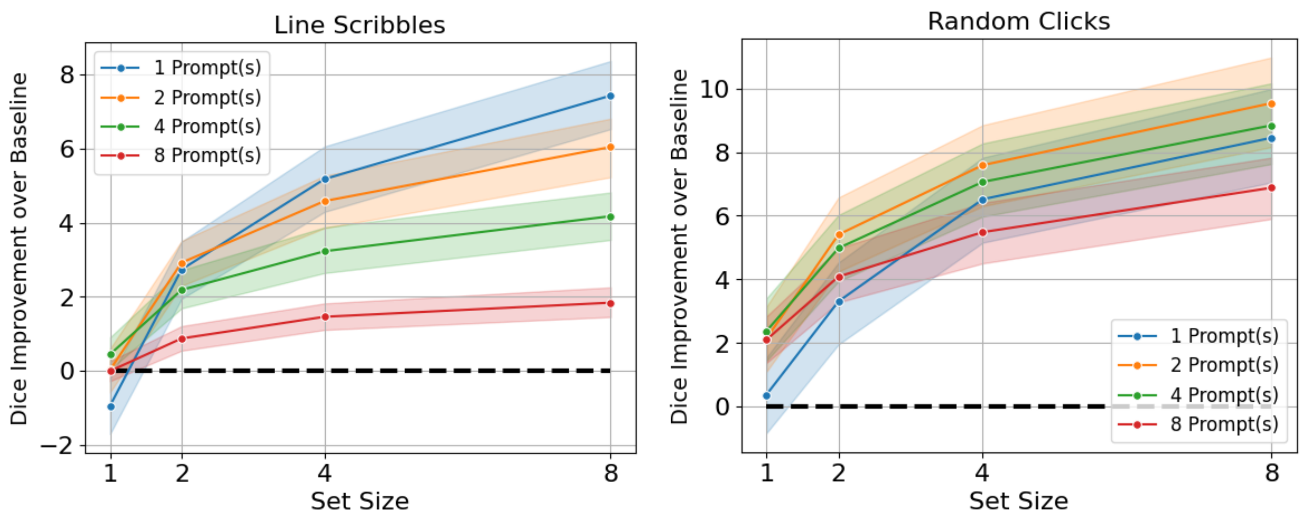
Discussion. Sample communication is especially useful in interactive segmentation because user-provided annotations on one image from the same task can provide information about the region being segmented on other images. Fig. 5 show that the baseline predictions, which only consider one sample at a time, are localized to the annotated pixels. By considering four annotated samples at once, ScribblePrompt-InterConv is able to gather more information about the target segmentation map from the whole set and propagate this information to each sample, achieving higher dice scores across all four predictions.

That the InterConv improvement over the baseline increases as the set size increases, demonstrates the value of shared information among samples. InterConv also achieves larger improvement with fewer prompts per image, and when there is less information per prompt (clicks as opposed to scribbles). In both situations, the less information there is per image, the more useful the information is from other images.

4.2. Automatic Segmentation: Clinical-quality MRI Scans

In this experiment, we evaluate InterConv for automatic segmentation of brain MRIs. MR images acquired in clinical settings often exhibit lower resolution, more noise, and lower tissue contrast than the research scans that are often used to evaluate segmentation methods. We evaluate InterConv on the challenging task of segmenting clinical brain MRI.

Data. *OASIS-3*: OASIS-3 is a collection of 1,378 partic-



(a) Interactive segmentation on Pandental dataset with line scribble prompts. (b) Interactive segmentation on ACDC dataset with line scribble prompts.

Figure 4. **ScribblePrompt-InterConv substantially improves interactive segmentation accuracy.** We evaluate ScribblePrompt-UNet and comparable-sized ScribblePrompt-InterConv on unseen data. Each plot shows shows mean per-sample Dice score *improvement* of ScribblePrompt-InterConv over baseline on various set sizes and prompt sizes. Shaded regions show 95% CI from bootstrapping with 1,000 runs.

	SimpleUNet	SimpleUNet-InterConv (ours)	SwinUNet	SwinUNet-InterConv (ours)	nnUNet	nnUNet-InterConv (ours)
OASIS-3 T2$\star$	83.44 ± 3.26	$\mathbf{83.65 \pm 3.39}$	81.10 ± 2.88	$\mathbf{81.78 \pm 3.06}$	82.62 ± 2.86	$\mathbf{83.50 \pm 2.98}$
OASIS-3 T2w-tse	90.28 ± 1.88	$\mathbf{90.71 \pm 1.60}$	87.58 ± 2.40	$\mathbf{88.02 \pm 2.13}$	89.69 ± 1.98	$\mathbf{89.90 \pm 1.67}$
ADNI FLAIR	86.95 ± 2.33	$\mathbf{87.95 \pm 2.32}$	85.25 ± 2.45	$\mathbf{85.40 \pm 2.38}$	87.23 ± 2.29	$\mathbf{88.48 \pm 2.18}$

Table 1. **Quantitative results: Dice performance on automatic segmentation.** We show the best performance for each model architecture along with standard deviation across test images.

ipants collected across several ongoing projects [66], containing 2,842 MR sessions with multiple modalities.

We focus on T2$\star$ and T2-TSE MRI modalities, which have higher noise, lower contrast, and sparser slices than research-quality T1 scans. Many subjects have multiple scan sessions acquired over time. For each session, we extract T1w scans, as well as T2-TSE and T2$\star$ scans, if they exist. We use segmentations obtained from the paired high-quality T1w image as ground truth, and use InterConv to predict segmentation maps from the (aligned) low-quality T2 scans. This enables us to assess the value of sharing information in low-quality information, while having accurate ground truth targets.

ADNI: We additionally use a collection of 5117 MR longitudinal sessions gathered from 1061 participants in the ADNI dataset. For each session, we extract a high-quality T1w scan and T2-FLAIR scan. The FLAIR images are acquired with large slice separation (5 mm), and therefore exhibit low-resolution in the slice dimension. As above, we obtain segmentations on the T1w image, and use it as the ground truth label map for the aligned FLAIR scans.

Processing. We first perform skull-stripping using SynthStrip [46]. Following standard practice in population analyses, we rigidly align all scans to a common space using SynthMorph [41–44]. Specifically, we first align all modalities within a session to the T1 scan of that session. For each subject, we then align the T1 scan of each session to the T1 scan of the *first* session. Finally, we align the T1 scan of the first session to the Talairach space, and propagate all scans to this space through the predicted transformations.

For both OASIS-3 and ADNI data, we use the middle coronal slice from all registered volumes. We filter and keep 20 anatomical regions that are commonly present in all scans. We exclude scans with missing labels. Details about class labels can be found in the Appendix.

We consider an input set to contain all scans from the same subject.

Network Architectures. We evaluate two architectural designs: UNet-based and Transformer-based. We train both SimpleUNets and SimpleUNet-InterConvs with varying model sizes from $\sim$60K to $\sim$16M parameters, as well as an nnUNet, which is automatically configured to have

CVPR
#13795

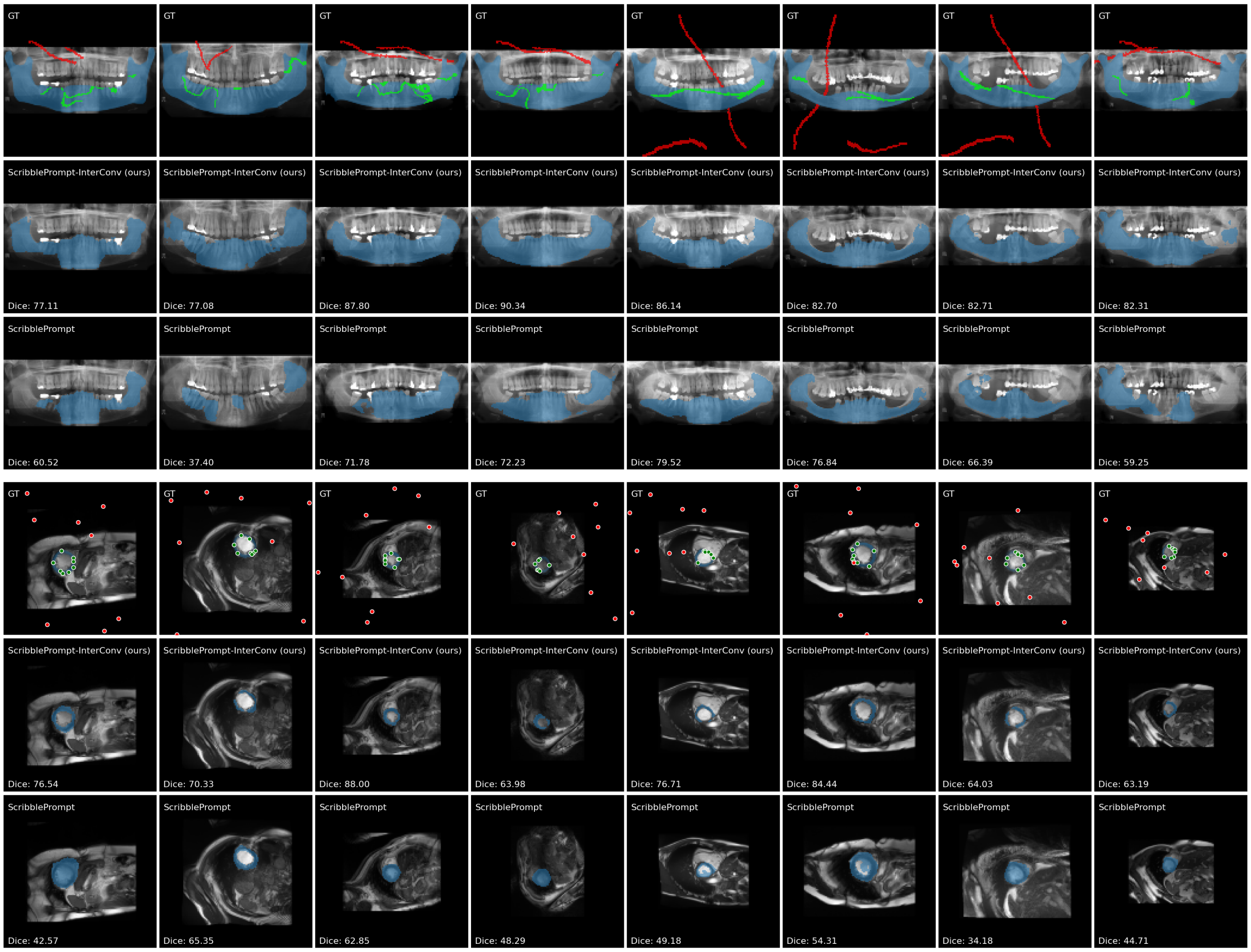


Figure 5. **ScribblePrompt-InterConv significantly improves interactive segmentation quality with limited prompts**. We run ScribblePrompt-InterConv and baseline predictions on 8 input images from Pandental, each provided with 2 positive (green) and 2 negative (red) line scribble prompts, and on 8 images from ACDC, providing each image with 8 positive (green) and 8 negative (red) random click prompts. The input samples come from different subjects within the same task. In each subfigure, **Top row** shows input image with provided prompts, along with ground truth segmentations, **middle row** shows ScribblePrompt-InterConv (ours) predictions, and **bottom row** shows ScribblePrompt (baseline) predictions overlayed in blue.

~20M parameters. We use the original SwinUNet design with ~26M parameters as a baseline for Transformer-based models.

We construct SimpleUNet-InterConv and nnUNet-InterConv by introducing a InterConv layer with kernel size 3 at the end of each convolution layer of SimpleUNets and nnUNet respectively. We construct SwinUNet-InterConv by adding a InterConv layer with kernel size 1 (i.e., point-wise feed-forward function) at the end of each multi-head attention layer of SwinUNet. Details about baseline network sizes and their corresponding InterConv-integrated network sizes can be found in the Appendix.

Training. *OASIS*: We train separate models for T2$\star$ and T2-TSE. For both T2$\star$ and T2wTSE, there are ~1030 subjects with the desired modality, which we first split into ~830 subjects with ~1350 images for model training, and ~100 held out test subjects with ~190 images. Details about train, validation and test data sizes can be found in the Appendix.

ADNI: With a total of 1061 subjects, we first split with 742 subjects for model training, and 213 subjects for performance evaluation.

We train all models with batches of 8 sets at each iteration, with a combination of Soft Dice Loss and Cross-entropy Loss:

$$\mathcal{L}_{seg}(\hat{y}, y) = \mathcal{L}_{\text{dice}}(\hat{y}, y) + \mathcal{L}_{\text{CCE}}(\hat{y}, y), \tag{6}$$

We train each model configuration of SimpleUNet and SwinUNet with 3 different weight initialization seeds and with

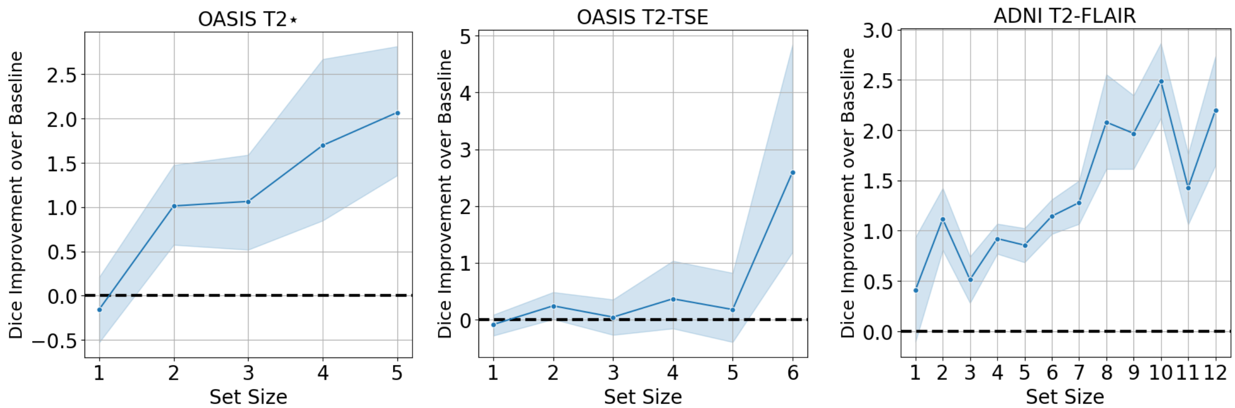


Figure 6. **nnUNet-InterConv achieves higher segmentation accuracy on subjects with more scans available.** We show performance improvement on nnUNet-InterConv over baseline nnUNet by averaging per-sample Dice improvement over subjects with a given number of scans. **Left** figure shows Dice improvement on OASIS T2⋆ scans. **Middle** figure shows Dice improvement on OASIS T2-TSE scans. **Right** figure shows Dice improvement on ADNI T2-FLAIR scans. Shaded regions show 95% CI from bootstrapping with 1,000 runs.

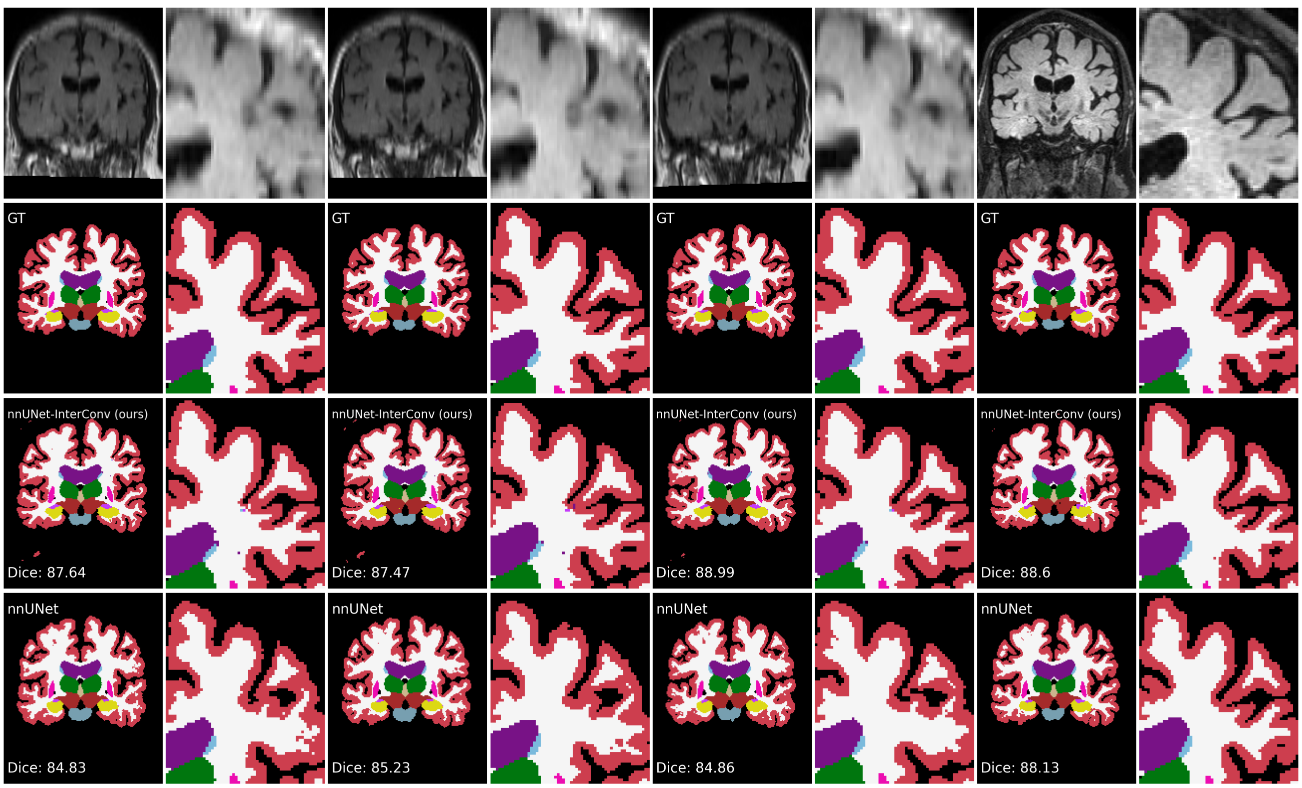


Figure 7. **nnUNet-InterConv generates substantially better prediction in the brain cortex with set interaction mechanism.** We visualize predictions on 4 randomly selected scans from an input subject with 8 scan sessions. We interleave inputs and segmentation maps with a close-up of the temporal lobe. **Top row**: input images, **Second row**: ground truth segmentation maps, **Third row**: nnuNet-InterConv predictions, **Fourth row**: baseline nnUNet predictions.

an early-stopping criterion based on validation loss not decreasing for 40,000 iterations. We initialize all nnUNet-structured models in the same way as the original nnUNet, and follow the original nnUNet training procedure to train the models for a fixed 1000 epochs with 250 iterations per epoch.

Evaluation. We analyze model performance as the average

Dice score across all test scan samples. For network configurations trained with multiple weight initializations, we average the predicted probabilities from each trained model before computing the Dice Score.

Results. Table 1 shows that all InterConv-integrated architectures improve segmentation accuracy over corresponding baselines. Fig. 6 shows that subjects with more scans available on average achieve better segmentation quality compared to those with fewer scans, demonstrating that larger sets are more informative. Fig. 7 shows that nnUNet-InterConv provides substantially better segmentations, e.g., of the cortex, than the baseline nnUNet. We analyze dice improvement and visualize results on the nnUNet framework since it consistently produces high segmentation accuracy and is a widely used baseline in the literature. Additional dice improvement analysis on other network architectures can be found in the Appendix.

Discussion. We emphasize that if the ground truth label maps are easily attainable from an individual image, we do not expect shared information to be helpful. To be able to even assess the value of shared information, we choose data where ground truth segmentation cannot be easily obtained from the scan itself, like the clinical-quality scans, and use the aligned research-quality scans to extract the ground truth maps.

The results in this section emphasize this effect: while InterConv outperforms corresponding baselines in all datasets, its improvement in OASIS T2-TSE is minimal, where image quality is least affected. In contrast, for the substantially more degraded T2* and FLAIR modalities, InterConv achieves substantial improvement over the baselines, by providing each image with additional shared information through set interaction.

5. Conclusion

We introduce InterConv, a mechanism that improves segmentation accuracy by enabling an existing segmentation model to benefit from jointly segmentating a set of related images. InterConv leverages intermediate features of each input sample to share information among the input set.

We demonstrate the value of this interaction by showing that InterConv-integrated networks outperforms baselines with the same architectural designs, in both automatic and interactive segmentation settings. For interactive segmentation, InterConv can reduce the total amount of labor required, since an annotation made on one image provides information that is useful for segmenting other images. For automatic segmentation, since existing segmentation models already perform well when the structures in an image are clearly defined, we focus on domains where individual images are challenging to segment, such as clinical-quality scans. We find that in these common settings, InterConv leads to improved segmentation quality.

The proposed InterConv Layer is easy to add to existing segmentation architectures and promises to help improve image segmentation in a wide variety of practical settings.

References

[1] Amir Hossein Abdi, Shohreh Kasaei, and Mojdeh Mehdizadeh. Automatic segmentation of mandible in panoramic x-ray. *Journal of Medical Imaging*, 2(4):044003, 2015. 2, 4, 1, 5

[2] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. 4, 1, 5

[3] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 6

[4] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4 (1):1–13, 2017. 6

[5] Sophia Bano, Francisco Vasconcelos, Luke M Shepherd, Emmanuel Vander Poorten, Tom Vercauteren, Sebastien Ourselin, Anna L David, Jan Deprest, and Danail Stoyanov. Deep placental vessel segmentation for fetoscopic mosaicking. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 763–773. Springer, 2020. 6

[6] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 119–127. Springer, 2019. 2

[7] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 4, 1, 5

[8] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 6

[9] Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution with-

out retraining. *Medical image analysis*, 86:102789, 2023. 2

[10] Nicholas Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Larry Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370(6):5, 2015. 6

[11] Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, 109: 218–225, 2019. 6

[12] Victor Ion Butoi*, Jose Javier Gonzalez Ortiz*, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. *International Conference on Computer Vision*, 2023. 2, 4, 1

[13] Juan C. Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Rohban, Shantanu Singh, and Anne E. Carpenter. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature Methods*, 16(12):1247–1253, 2019. 6

[14] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision Workshops(ECCVW)*, 2022. 3

[15] Albert Cardona, Stephan Saalfeld, Stephan Preibisch, Benjamin Schmid, Anchi Cheng, Jim Pulokas, Pavel Tomancak, and Volker Hartenstein. An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. *PLoS biology*, 8(10):e1000502, 2010. 6

[16] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[17] Lele Chen, Yue Wu, Adora M DSouza, Anas Z Abidin, Axel Wismüller, and Chenliang Xu. Mri tumor segmentation with densely connected 3d cnn. In *Medical imaging 2018: image processing*, pages 357–364. SPIE, 2018. 2

[18] Albert Clèrigues, Sergi Valverde, Jose Bernal, Jordi Freixenet, Arnau Oliver, and Xavier Lladó. Acute and sub-acute stroke lesion segmentation from multimodal mri. *Computer methods and programs in biomedicine*, 194:105521, 2020. 2

[19] Noel C. F. Codella, David A. Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1710.05006, 2017. 6

[20] Shaoguo Cui, Lei Mao, Jingfeng Jiang, Chang Liu, Shuyu Xiong, et al. Automatic semantic segmentation of brain gliomas from mri images using a deep cascaded neural network. *Journal of healthcare engineering*, 2018, 2018. 2

[21] Steffen Czolbe and Adrian V Dalca. Neuralizer: General neuroimage analysis without re-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6217–6230, 2023. 2

[22] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018. 6

[23] Adrian V Dalca, Evan Yu, Polina Golland, Bruce Fischl, Mert R Sabuncu, and Juan Eugenio Iglesias. Unsupervised deep learning for bayesian brain mri segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 356–365. Springer, 2019. 2

[24] Etienne Decenciere, Guy Cazuguel, Xiwei Zhang, Guillaume Thibault, J-C Klein, Fernand Meyer, Beatriz Marcotegui, Gwénolé Quellec, Mathieu Lamard, Ronan Danno, et al. Teleophta: Machine learning and image processing methods for teleophthalmology. *Irbm*, 34(2):196–203, 2013. 6

[25] Aysen Degerli, Morteza Zabihi, Serkan Kiranyaz, Tahir Hamid, Rashid Mazhar, Ridha Hamila, and Moncef Gabbouj. Early detection of myocardial infarction in low-quality echocardiography. *IEEE Access*, 9:34442–34453, 2021. 6

[26] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 3, 4

[27] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. Ivdnet: Intervertebral disc localization and segmentation in mri with a multi-modal unet. In *International workshop and challenge on computational methods and clinical applications for spine imaging*, pages 130–143. Springer, 2018. 2

[28] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. *IEEE transactions on medical imaging*, 38 (5):1116–1126, 2018. 2

[29] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38 (5):1116–1126, 2019. 2

[30] J Gamper, NA Koohbanani, K Benes, S Graham, M Jahanifar, SA Khurram, A Azam, K Hewitt, and N Rajpoot. Pannuke dataset extension, insights and baselines. arxiv. 2020 doi: 10.48550. *ARXIV*, 2003. 6

[31] Stephan Gerhard, Jan Funke, Julien Martel, Albert Cardona, and Richard Fetter. Segmented anisotropic ssTEM dataset of neural tissue. 2013. 6

[32] Randy L Gollub, Jody M Shoemaker, Margaret D King, Tonya White, Stefan Ehrlich, Scott R Sponheim, Vincent P

Clark, Jessica A Turner, Bryon A Mueller, Vince Magnotta, et al. The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 11:367–388, 2013. 6

[33] Ioannis S Gousias, Daniel Rueckert, Rolf A Heckemann, Leigh E Dyet, James P Boardman, A David Edwards, and Alexander Hammers. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. *Neuroimage*, 40(2):672–684, 2008. 6

[34] Ioannis S Gousias, A David Edwards, Mary A Rutherford, Serena J Counsell, Jo V Hajnal, Daniel Rueckert, and Alexander Hammers. Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. *Neuroimage*, 62(3):1499–1509, 2012. 6

[35] Endre Grøvik, Darvin Yi, Michael Iv, Elizabeth Tong, Daniel Rubin, and Greg Zaharchuk. Deep learning enables automatic detection and segmentation of brain metastases on multisequence mri. *Journal of Magnetic Resonance Imaging*, 51(1):175–182, 2020. 6

[36] Daniel Gut. X-ray images of the hip joints. 1, 2021. Publisher: Mendeley Data. 4, 1, 5

[37] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 2

[38] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, page 101821, 2020. 6

[39] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1):762, 2022. 6

[40] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1):762, 2022. 2

[41] Malte Hoffmann, Benjamin Billot, Juan E Iglesias, Bruce Fischl, and Adrian V Dalca. Learning mri contrast-agnostic registration. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 899–903. IEEE, 2021. 5

[42] Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE Transactions on Medical Imaging*, 41(3):543–558, 2022.

[43] Malte Hoffmann, Andrew Hoopes, Bruce Fischl, and Adrian V Dalca. Anatomy-specific acquisition-agnostic affine registration learned from fictitious images. In *Medical Imaging 2023: Image Processing*, page 1246402. SPIE, 2023.

[44] Malte Hoffmann, Andrew Hoopes, Douglas N Greve, Bruce Fischl, and Adrian V Dalca. Anatomy-aware and acquisition-agnostic joint registration with synthmorph. *arXiv preprint arXiv:2301.11329*, 2023. 5

[45] Andrew Hoopes, Malte Hoffmann, Douglas N. Greve, Bruce Fischl, John Guttag, and Adrian V. Dalca. Learning the effect of registration hyperparameters with hypermorph. *Machine Learning for Biomedical Imaging*, 1:1–30, 2022. 2, 6

[46] Andrew Hoopes, Jocelyn S Mora, Adrian V Dalca, Bruce Fischl, and Malte Hoffmann. Synthstrip: skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022. 5

[47] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000. 6

[48] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 2

[49] Humans in the Loop. Teeth segmentation dataset. 6

[50] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*, pages 287–297. Springer, 2018. 2

[51] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 2, 3

[52] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. *arXiv preprint arXiv:2404.09556*, 2024. 2, 3

[53] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. 6

[54] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017. 2

[55] Rashed Karim, R James Housden, Mayuragoban Balasubramaniam, Zhong Chen, Daniel Perry, Ayesha Uddin, Yosra

Al-Beyatti, Ebrahim Palkhi, Prince Acheampong, Samantha Obom, et al. Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. *Journal of Cardiovascular Magnetic Resonance*, 15(1):1–17, 2013. 6

[56] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, 2019. 6

[57] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 2, 6

[58] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[59] Serkan Kiranyaz, Aysen Degerli, Tahir Hamid, Rashid Mazhar, Rayyan El Fadil Ahmed, Rayaan Abouhasera, Morteza Zabihi, Junaid Malik, Ridha Hamila, and Moncef Gabbouj. Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. *IEEE Access*, 8:210301–210317, 2020. 6

[60] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2

[61] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018. 2

[62] Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, Lemonia Konstantinidou, Marcus Makowski, James Nagarajah, Nassir Navab, et al. Tracked 3d ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *Plos one*, 17(7):e0268550, 2022. 6

[63] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019. 6

[64] Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J Counsell, Valentina Doria, Ahmed Serag, Ioannis S Gousias, James P Boardman, Mary A Rutherford, A David Edwards, et al. A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011. 6

[65] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2020. 6

[66] PJ LaMontague, TLS Benzinger, John C Morris, S Keefe, R Hornbeck, C Xiong, E Grant, J Hassenstab, K Moulder, AG Vlassenko, et al. Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medRxiv*, page 2019, 2019. 5

[67] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault Workshop Challenge*, page 12, 2015. 4, 1, 5, 6

[68] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019. 6

[69] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015. 6

[70] Mingchao Li, Yuhan Zhang, Zexuan Ji, Keren Xie, Songtao Yuan, Qinghuai Liu, and Qiang Chen. Ipn-v2 and octa-500: Methodology and dataset for retinal image segmentation. *arXiv preprint arXiv:2012.07261*, 2020. 6

[71] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[72] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 6

[73] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2

[74] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012. 6

[75] Maximilian T Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020. 6

[76] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021. 6

[77] Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82: 102616, 2022. 6

[78] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 2

[79] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 2

[80] Yuhui Ma, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE Transactions on Medical Imaging*, 40(3):928–939, 2021. 6

[81] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9): 1498–1507, 2007. 2, 6

[82] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011. 6

[83] Francesco Marzola, Nens Van Alfen, Jonne Doorduin, and Kristen M. Meiburger. Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. *Computers in Biology and Medicine*, 135:104623, 2021. 6

[84] Maciej A Mazurowski, Kal Clark, Nicholas M Czarnek, Parisa Shamsesfandabadi, Katherine B Peters, and Ashirbani Saha. Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data. *Journal of neuro-oncology*, 133:27–35, 2017. 6

[85] Bjoern Menze, Leo Joskowicz, Spyridon Bakas, Andras Jakab, Ender Konukoglu, Anton Becker, Amber Simpson, and Richard D. Quantification of uncertainties in biomedical image quantification 2021. *4th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021)*, 2021. 6

[86] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 2, 6

[87] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 4

[88] Anna Montoya, Hasnin, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound nerve segmentation, 2016. 6

[89] Kelly Payette, Priscille de Dumast, Hamza Kebiri, Ivan Ezhov, Johannes C Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, et al. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific Data*, 8(1):1–14, 2021. 6

[90] Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th international symposium on medical information processing and analysis*, page 92870W. SPIE / International Society for Optics and Photonics, 2015. 6

[91] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016. 2

[92] Gašper Podobnik, Primož Strojan, Primož Peterlin, Bulat Ibragimov, and Tomaž Vrtovec. HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset. *Medical Physics*, 50(3):1917–1927, 2023. tex.eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.16197 6

[93] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018. 6

[94] Perry Radau, Yingli Lu, Kim Connelly, Gideon Paul, AJWG Dick, and Graham Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, 49, 2009. 4, 1, 5

[95] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11536–11546, 2023. 2

[96] Marianne Rakic, Hallee E. Wong, Jose Javier Gonzalez Ortiz, Beth Cimini, John V. Guttag, and Adrian V. Dalca. Tyche: Stochastic in-context learning for medical image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2, 4, 1

[97] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L. Rubin. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020. 6

[98] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[99] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, SQ Truong, CD Nguyen, Van-Doan Ngo, Jayne

Seekins, Francis G Blankenberg, AY Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *MedRxiv*, 2021. 6

[100] Constantin Seibold, Simon Reiß, Saquib Sarfraz, Matthias A. Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H. Maier-Hein, Jens Kleesiek, and Rainer Stiefelhagen. Detailed annotations of chest x-rays via ct projection for report understanding. In *Proceedings of the 33th British Machine Vision Conference (BMVC)*, 2022. 6

[101] Ahmed Serag, Paul Aljabar, Gareth Ball, Serena J Counsell, James P Boardman, Mary A Rutherford, A David Edwards, Joseph V Hajnal, and Daniel Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *Neuroimage*, 59(3):2255–2265, 2012. 6

[102] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017. 6

[103] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1214–1223, 2021. 2

[104] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 6

[105] Samuel L Smith, Andrew Brock, Leonard Berrada, and Soham De. Convnets match vision transformers at scale. *arXiv preprint arXiv:2310.16764*, 2023. 2

[106] Yuxin Song, Jing Zheng, Long Lei, Zhipeng Ni, Baoliang Zhao, and Ying Hu. CT2US: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics*, 122:106706, 2022. 6

[107] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004. 2, 4, 1, 5

[108] Vajira Thambawita, Steven A Hicks, Pål Halvorsen, and Michael A Riegler. Divergentnets: Medical image segmentation by network ensemble. *arXiv preprint arXiv:2107.00283*, 2021. 2

[109] Santiago Vitale, José Ignacio Orlando, Emmanuel Iarussi, and Ignacio Larrabide. Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. *International journal of computer assisted radiology and surgery*, 15(2):183–192, 2020. 6

[110] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019. 2

[111] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 2

[112] Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024. 2

[113] Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. Scribbleprompt: Fast and flexible interactive segmentation for any biomedical image. *arXiv:2312.07381*, 2024. 2, 3, 4, 1, 6

[114] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 2

[115] Lukas Zbinden, Lars Doorenbos, Theodoros Pissas, Adrian Thomas Huber, Raphael Sznitman, and Pablo Márquez-Neila. Stochastic segmentation with conditional categorical diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1119–1129, 2023. 2

[116] Yingtao Zhang, Min Xian, Heng-Da Cheng, Bryar Shareef, Jianrui Ding, Fei Xu, Kuan Huang, Boyu Zhang, Chunping Ning, and Ying Wang. Busis: A benchmark for breast ultrasound image segmentation. In *Healthcare*, page 729. MDPI, 2022. 6

[117] Qi Zhao, Shuchang Lyu, Wenpei Bai, Linghan Cai, Binghao Liu, Meijing Wu, Xiubo Sang, Min Yang, and Lijiang Chen. A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. *CoRR*, abs/2207.06799, 2022. 6

[118] Xiaomei Zhao, Yihong Wu, Guidong Song, Zhenye Li, Yazhuo Zhang, and Yong Fan. A deep learning model integrating fcnns and crfs for brain tumor segmentation. *Medical image analysis*, 43:98–111, 2018. 2

[119] Guoyan Zheng, Chengwen Chu, Daniel L Belavỳ, Bulat Ibragimov, Robert Korez, Tomaž Vrtovec, Hugo Hutt, Richard Everson, Judith Meakin, Isabel Lŏpez Andrade, et al. Evaluation and comparison of 3d intervertebral disc localization and segmentation methods for 3d t2 mr data: A grand challenge. *Medical image analysis*, 35:327–344, 2017. 4, 1, 5

[120] Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018. 2, 4, 1, 5

CVPR
#13795

InterConv: Set Interaction for Improved Biomedical Image Segmentation

Supplementary Material

Label	Class
0	Background
1	Left cerebral white matter
2	Left cerebral cortex
3	Left lateral ventricle
4	Left inferior lateral ventricle
5	Left thalamus
6	Left caudate
7	Left putamen
8	3rd Ventricle
9	Brain stem
10	Left hippocampus
11	Left ventral DC
12	Right cerebral white matter
13	Right cerebral cortex
14	Right lateral ventricle
15	Right inferior lateral ventricle
16	Right thalamus
17	Right caudate
18	Right putamen
19	Right hippocampus
20	Right ventral DC

Table 2. **OASIS and ADNI class labels**.

6. Automatic Segmentation

6.1. Data

We provide training, validation and test sample sizes in Table 3. Each dataset is split into four subsets:

- Train: during training stage used by the model for gradient descent.
- Stopping-Val: during training stage used for early stopping.
- Performance-Val: during performance evaluation stage used for choosing the best model configuration.
- Test: used for reporting model performance.

Set size distribution in each of the data split is shown in Fig. 8.

We perform the experiment in 2D since we thoroughly evaluate many variants of our method and the baseline, which would be prohibitive in 3D. After pre-processing, each scan volume is of size: (160, 192, 224). We take the 109th coronal slice. The resulting slices are of size 160×192. We list the class labels in Table 2.

6.2. Model Configuration

We apply InterConv to three architectural designs: SimpleUNet, nnUNet, and SwinUNet.

SimpleUNets have 4 encoder layers with the same feature size at each layer. We train baselines and SimpleUNet-InterConv with comparable number of parameters. We detail various network sizes and corresponding output feature sizes per layer in Table 4.

We designed nnUNet-InterConv and SwinUNet-InterConv using the same network backbone, with both the same output feature size per layer as the original model, and with adjusted feature sizes to align InterConv-integrated models with baselines. We include detailed layer-wise feature configurations and their corresponding network sizes in Table 5 and 6.

6.3. Additional Experiments

Fig. 9 and 10 show that larger set size enables sharing of more informative accross the set for the SimpleUNet and SwinUNet architectures as well (along with the nnUNet shown in the main paper). Subjects with more scans available on average achieve better segmentation quality compared to those with fewer scans.

7. Interactive Segmentation

7.1. Data

Datasets. Building upon large data collection efforts like MegaMedical [12, 96, 113], we use a collection of 74 biomedical image segmentation datasets. We divide the collection into 65 training datasets (Table 8) and 9 evaluation datasets (Table 7) following the same partitions as [113]. The evaluation datasets include a diverse array of biomedical domains including eyes [107], abdominal [67], cardiac [7, 94], bones [36], teeth [1], spine [119], cells [120], and lesions [2].

Each dataset was split into 60% training, 20% validation and 20% test as in [12, 96, 113], although not all splits were used. Each model was trained on the training splits of the training datasets. We report final results on the test split of the evaluation datasets.

Tasks. As in [12, 96, 113], we define a 2D segmentation task as a combination of (sub)dataset, axis (for 3D modalities), and label. For datasets with sub-datasets (e.g., malignant vs. benign lesions), each cohort is considered a separate task. Each segmentation label in multi-label datasets

	Subjects	Images
Train	777	1332
Stopping-Val	50	82
Performance-Val	47	82
Test	102	187

(a) OASIS T2⋆

	Subjects	Images
Train	780	1381
Stopping-Val	49	81
Performance-Val	48	87
Test	99	188

(b) OASIS T2-TSE

	Subjects	Images
Train	636	3075
Stopping-Val	106	493
Performance-Val	106	504
Test	213	1045

(c) ADNI T2-FLAIR

Table 3. **Sample Sizes** for automatic segmentation experiments.

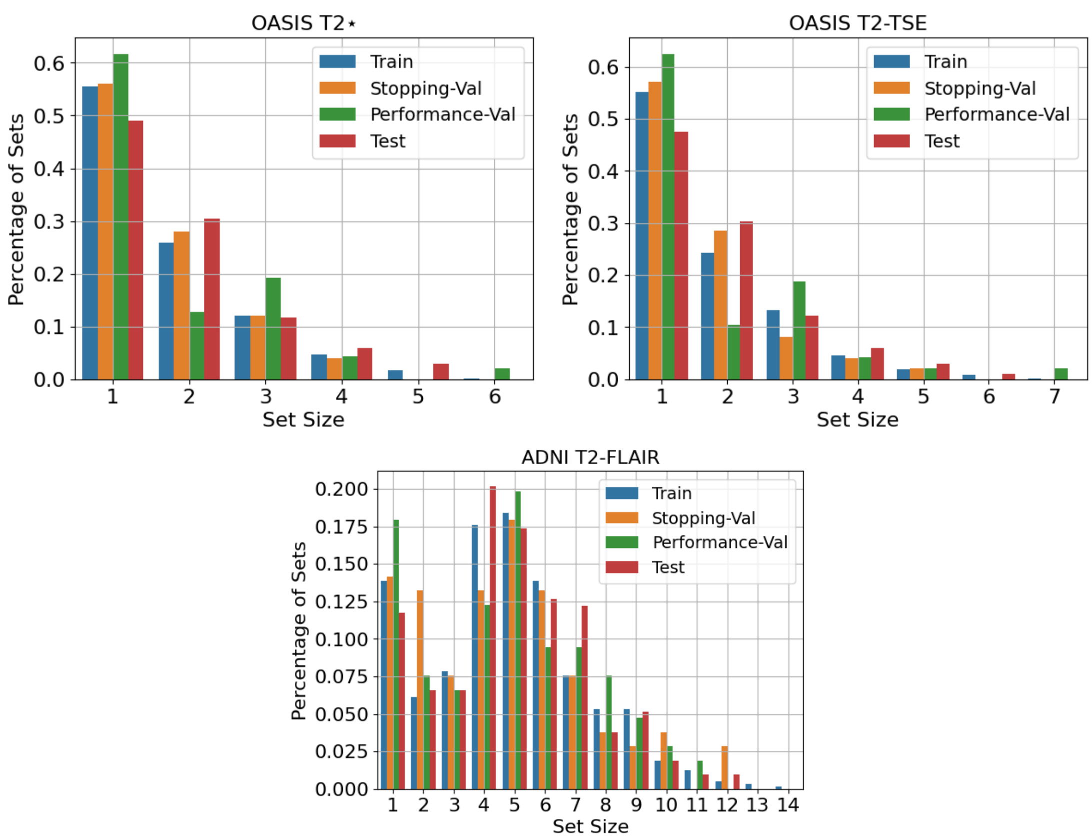


Figure 8. **Set Size distribution within splits for OASIS and ADNI data.** We show the various number of scans for train, validation and test subjects, as well as the distribution in the test, stopping-val, performance-val and test splits.

is treated as a separate binary task. For multi-annotator datasets, each annotator is considered a separate label. In instance segmentation datasets, we train on one instance at a time. For 3D modalities, we use the slice with the maximum label area (”maxslice”) for each subject.

Image Processing. We resize images to 128^2 and rescale image intensities to [0,1].

	# ouput features per layer	
Network Size (# of parameters)	**SimpleUNet-InterConv**	**Baseline**
60K	16	24
120K	23	34
240K	33	48
500K	49	70
1M	70	100
2M	100	142
4M	141	200
8M	200	284
16M	282	400

Table 4. Network sizes for SimpleUNet architectures.

	Encoder Feature Dimensions	**Network Size**
nnUNet	$[32, 64, 128, 256, 512, 512]$	20.6M
nnUNet-InterConv	$[23, 46, 92, 184, 368, 368]$	20.4M
	$[32, 64, 128, 256, 512, 512]$	34.9M

Table 5. Network sizes for nnUNet architectures.

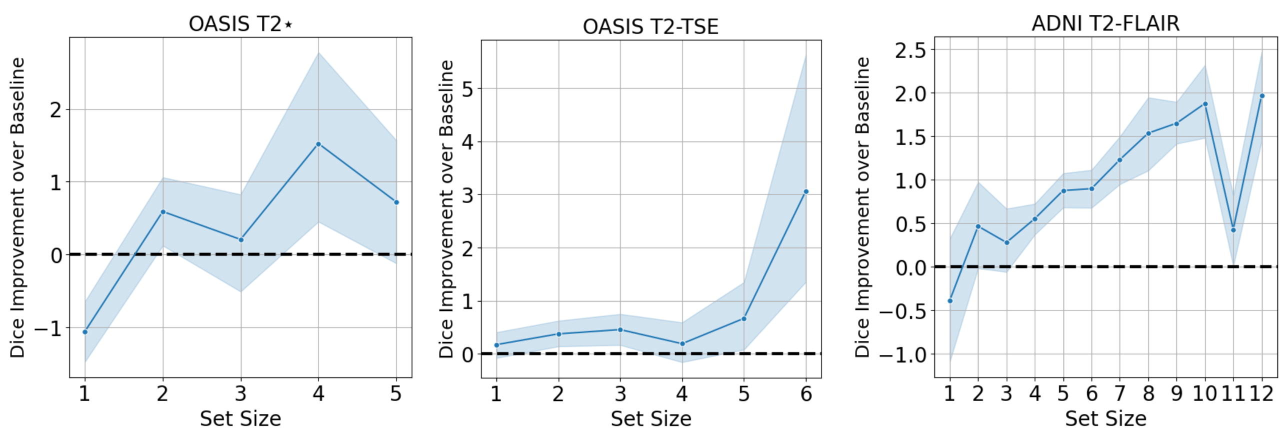


Figure 9. **SimpleUNet-InterConv achieves higher segmentation accuracy on subjects with more scans available.** We show performance improvement on SimpleUNet-InterConv over baseline SimpleUNet by averaging per-sample Dice improvement over subjects with a given number of scans. We evaluate with models with the best average test performance. **Left** figure shows Dice improvement on OASIS T2⋆ scans. **Middle** figure shows Dice improvement on OASIS T2-TSE scans. **Right** figure shows Dice improvement on ADNI T2-FLAIR scans. Shaded regions show 95% CI from bootstrapping with 1,000 runs.

7.2. Additional Experiments

We provide additional interactive segmentation visualizations with random click prompts in Fig. 11. ScribblePrompt-InterConv is able to leverage the information across the input set to substantially improve the segmentations compared to the baseline models, especially when given few user interactions.

	Embedding Dimension	Network Size
SwinUNet	96	27.1M
SwinUNet-InterConv	84	26.8M
	96	34.9M

Table 6. Network sizes for SwinUNet architectures.

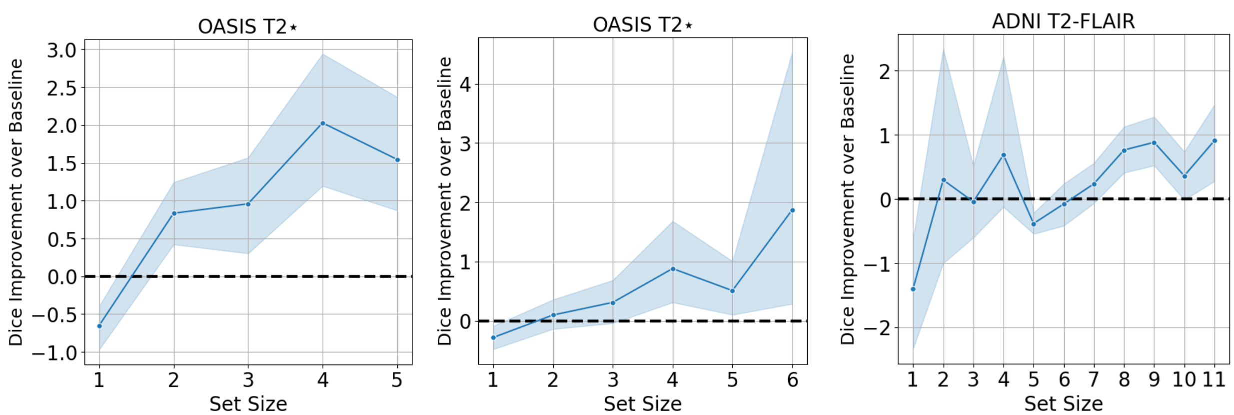
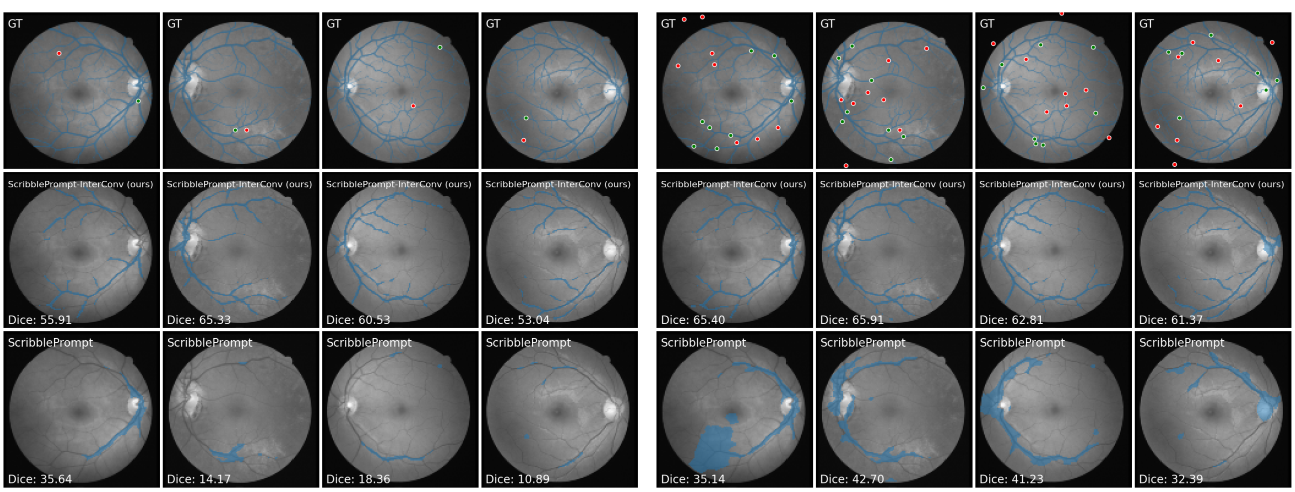


Figure 10. **SwinUNet-InterConv generally achieves higher segmentation accuracy on subjects with more scans available.** We show performance improvement on SwinUNet-InterConv over baseline SwinUNet by averaging per-sample Dice improvement over subjects with a given number of scans. We evaluate with models with the best average test performance. **Left** figure shows Dice improvement on OASIS T2⋆ scans. **Middle** figure shows Dice improvement on OASIS T2-TSE scans. **Right** figure shows Dice improvement on ADNI T2-FLAIR scans. Shaded regions show 95% CI from bootstrapping with 1,000 runs.



(a) Segmentation visualization with 1 positive and 1 negative prompt

(b) Segmentation visualization with 8 positive and 8 negative prompts

Figure 11. **ScribblePrompt-InterConv significantly improves interactive segmentation quality with random click prompts**. We run ScribblePrompt-InterConv and baseline predictions on 4 input images from DRIVE, providing each image with 1 positive (green) and 1 negative (red) random click prompt, and on the same set of images with 8 positive (green) and 8 negative (red) random click prompts. The input samples come from different subjects within the same task. In each subfigure, **Top row** shows input image with provided prompts, along with ground truth segmentations, **middle row** shows ScribblePrompt-InterConv (ours) predictions, and **bottom row** shows ScribblePrompt (baseline) predictions overlayed in blue.

Dataset Name	Description	Scans	Labels	Modalities
ACDC [7]	Left and right ventricular endocardium	99	3	cine-MRI
BTCV Cervix [67]	Bladder, uterus, rectum, small bowel	30	4	CT
BUID [2]	Breast tumors	647	2	Ultrasound
DRIVE [107]	Blood vessels in retinal images	20	1	Optical camera
HipXRay [36]	Ilium and femur	140	2	X-Ray
PanDental [1]	Mandible and teeth	215	2	X-Ray
SCD [94]	Sunnybrook Cardiac Multi-Dataset Collection	100	1	cine-MRI
SpineWeb [119]	Vertebrae	15	1	T2-weighted MRI
WBC [120]	White blood cell cytoplasm and nucleus	400	2	Microscopy

Table 7. **Evaluation datasets**. For the relative size of datasets, we include the number of unique scans (subject and modality pairs) that each dataset has. These datasets were unseen by the models during training. The validation splits of the datasets were used for model selection. We report final results on the test splits of these datasets.

Dataset Name	Description	Scans	Modalities
AbdominalUS [109]	Abdominal organ segmentation	1,543	Ultrasound
AMOS [53]	Abdominal organ segmentation	240	CT, MRI
BBBC003 [74]	Mouse embryos	15	Microscopy
BBBC038 [13]	Nuclei instance segmentation	670	Microscopy
BrainDev [33, 34, 64, 101]	Adult and neonatal brain atlases	53	Multimodal MRI
BrainMetShare[35]	Brain tumors	420	Multimodal MRI
BRATS [3, 4, 86]	Brain tumors	6,096	Multimodal MRI
BTCV Abdominal [67]	13 abdominal organs	30	CT
BUSIS [116]	Breast tumors	163	Ultrasound
CAMUS [68]	Four-chamber and Apical two-chamber heart	500	Ultrasound
CDemris [55]	Human left atrial wall	60	CMR
CHAOS [56, 57]	Abdominal organs (liver, kidneys, spleen)	40	CT, T2-weighted MRI
CheXplanation [99]	Chest X-Ray observations	170	X-Ray
CT2US [106]	Liver segmentation in synthetic ultrasound	4,586	Ultrasound
CT-ORG[97]	Abdominal organ segmentation (overlap with LiTS)	140	CT
DDTI [90]	Thyroid segmentation	472	Ultrasound
EOphtha [24]	Eye microaneurysms and diabetic retinopathy	102	Optical camera
FeTA [89]	Fetal brain structures	80	Fetal MRI
FetoPlac [5]	Placenta vessel	6	Fetoscopic optical camera
FLARE [77]	Abdominal organs (liver, kidney, spleen, pancreas)	361	CT
HaN-Seg [92]	Head and neck organs at risk	84	CT, T1-weighted MRI
HMC-QU [25, 59]	4-chamber (A4C) and apical 2-chamber (A2C) left wall	292	Ultrasound
I2CVB [69]	Prostate (peripheral zone, central gland)	19	T2-weighted MRI
IDRID [93]	Diabetic retinopathy	54	Optical camera
ISBI-EM [15]	Neuronal structures in electron microscopy	30	Microscopy
ISIC [19]	Demoscopic lesions	2,000	Dermatology
ISLES [39]	Ischemic stroke lesion	180	Multimodal MRI
KiTS [38]	Kidney and kidney tumor	210	CT
LGGFlair [11, 84]	TCIA lower-grade glioma brain tumor	110	MRI
LiTS [8]	Liver tumor	131	CT
LUNA [102]	Lungs	888	CT
MCIC [32]	Multi-site brain regions of schizophrenic patients	390	T1-weighted MRI
MMOTU [117]	Ovarian tumors	1,140	Ultrasound
MSD [104]	Large-scale collection of 10 medical segmentation datasets	3,225	CT, Multimodal MRI
MuscleUS [83]	Muscle segmentation (biceps and lower leg)	8,169	Ultrasound
NCI-ISBI [10]	Prostate	30	T2-weighted MRI
NerveUS [88]	Nerve segmentation	5,635	Ultrasound
OASIS [45, 81]	Brain anatomy	414	T1-weighted MRI
OCTA500 [70]	Retinal vascular	500	OCT/OCTA
PanNuke [30]	Nuclei instance segmentation	7,901	Microscopy
PAXRay [100]	92 labels covering lungs, mediastinum, bones, and sub-diaphram in Chest X-Ray	852	X-Ray
PROMISE12 [72]	Prostate	37	T2-weighted MRI
PPMI [22, 82]	Brain regions of Parkinson patients	1,130	T1-weighted MRI
QUBIQ [85]	Collection of 4 multi-annotator datasets (brain, kidney, pancreas and prostate)	209	T1-weighted MRI, Multimodal MRI, CT
ROSE [80]	Retinal vessel	117	OCT/OCTA
SegTHOR [65]	Thoracic organs (heart, trachea, esophagus)	40	CT
SegThy [62]	Thyroid and neck segmentation	532	MRI, Ultrasound
ssTEM [31]	Neuron membranes, mitochondria, synapses and extracellular space	20	Microscopy
STARE [47]	Blood vessels in retinal images (multi-annotator)	20	Optical camera
ToothSeg [49]	Individual teeth	598	X-Ray
VerSe [75]	Individual vertebrae	55	CT
WMH [63]	White matter hyper-intensities	60	Multimodal MRI
WORD [76]	Abdominal organ segmentation	120	CT

Table 8. **Training datasets**. We use the same set of 65 training datasets as [113]. For the relative size of datasets, we show the number of unique scans (subject and modality pairs) that each dataset has.