Adapting Pre-trained Vision Transformers from 2D to 3D through Weight Inflation Improves Medical Image Segmentation

Yuhui Zhang   YUHUIZ@STANFORD.EDU
Shih-Cheng Huang   MSCHUANG@STANFORD.EDU
Zhengping Zhou   ZPZHOU@STANFORD.EDU
Matthew P. Lungren   MLUNGREN@STANFORD.EDU
Serena Yeung   SYYEUNG@STANFORD.EDU
Stanford University, Stanford, CA 94305, USA

Abstract

Given the prevalence of 3D medical imaging technologies such as MRI and CT that are widely used in diagnosing and treating diverse diseases, 3D segmentation is one of the fundamental tasks of medical image analysis. Recently, Transformer-based models have started to achieve state-of-the-art performances across many vision tasks, through pre-training on large-scale natural image benchmark datasets. While works on medical image analysis have also begun to explore Transformer-based models, there is currently no optimal strategy to effectively leverage pre-trained Transformers, primarily due to the difference in dimensionality between 2D natural images and 3D medical images. Existing solutions either split 3D images into 2D slices and predict each slice independently, thereby losing crucial depth-wise information, or modify the Transformer architecture to support 3D inputs without leveraging pre-trained weights. In this work, we use a simple yet effective weight inflation strategy to adapt pre-trained Transformers from 2D to 3D, retaining the benefit of both transfer learning and depth information. We further investigate the effectiveness of transfer from different pre-training sources and ob-jectives. Our approach achieves state-of-the-art performances across a broad range of 3D medical image datasets, and can become a standard strategy easily utilized by all work on Transformer-based models for 3D medical images, to maximize performance.[1]

Keywords: Medical Image Segmentation, Transfer Learning, CT, MRI.

1. Introduction

The increased utilization of medical imaging technologies in recent years is causing a dramatic increase in radiologists' daily workload (Smith-Bindman et al., 2008; Hendee et al., 2010), leading to longer delays in diagnosis and higher misdiagnosis rates (Alonso-Martínez et al., 2010; Hendriksen et al., 2017). Computer vision techniques such as 3D segmentation have the potential to alleviate the burden for radiologists by automating or assisting current image interpretation workflow, ultimately improving clinical care and patient outcome (Esteva et al., 2021; Huang et al., 2020).

Recently, the computer vision community has witnessed a paradigm shift from using CNNs to self-attention based Trans-

---

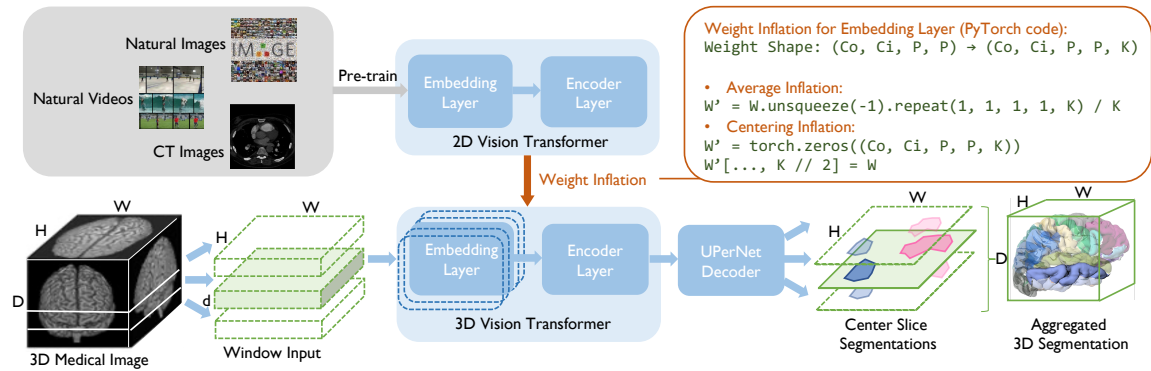1. Codes are available at https://github.com/yuhui-zh15/TransSeg.

Figure 1: *Approach overview.* Large-scale pre-trained Transformers are used as the encoder in the segmentation model for transfer learning, in which weights are adapted using the inflation strategy to support 3D inputs. Each 3D image is split into windows, which contain a small number of neighbor slices. Each window is fed into the segmentation model and the segmentation of the center slice is predicted. All the predicted slices are aggregated to form the final 3D prediction.

former (Vaswani et al., 2017) architectures. By pre-training at scale using supervised or self-supervised learning, many vision Transformer models have achieved state-of-the-art performances when fine-tuned for various vision tasks, including image classification, object detection, and semantic segmentation (Caron et al., 2021; Liu et al., 2021b; Bao et al., 2021; Carion et al., 2020; Dosovitskiy et al., 2020).

However, while many recent works on medical image analysis have started to explore Transformer-based models, it is still unclear what is the optimal strategy to effectively leverage Transformers for 3D medical image segmentation, primarily due to the difference in dimensionality between 2D natural images and 3D medical images. A straightforward approach adopted by most existing works is to split 3D images into 2D slices along the depth axis and independently segment each of them (Chen et al., 2021; Liu et al., 2021a; Huang et al., 2021). However, this solution compromises the depth information, which is crucial for identifying

segmentation boundaries. On the contrary, another line of research modifies the Transformer architecture to support 3D inputs, but the modification makes the pre-trained weights not directly applicable (Xie et al., 2021; Hatamizadeh et al., 2021).

In this work, we investigate strategies and best practices for adapting Transformers for 3D medical images. We first analyze the importance of *transfer learning* and *depth information*, and find that both are critical for segmentation performance, especially transfer learning. The model initialized from pre-trained weights outperforms the same randomly initialized model by 11.18% on a multi-organ segmentation dataset, and the randomly initialized 3D model outperforms the randomly initialized 2D model by 6.07%. To retain these two advantages, we use a simple yet effective *weight inflation* strategy to adapt pre-trained Transformers from 2D to 3D, which has been a standard approach in video understanding to transfer from models pre-trained on images since it was first proposed in I3D (Carreira and Zisserman,

392

2017). After careful ablations of different inflation and transfer settings, our best strategy achieves 1.95% improvements over baselines with only a 0.75% increase in computational cost, and outperforms all the state-of-the-art methods.

We further evaluate our method on 11 additional datasets to understand the generalizability of our best practice. Experiments show that our approach consistently benefits from weight inflation and achieves many state-of-the-art performances.

In summary, the major contribution of our work to the medical image segmentation community is that we show how a simple yet effective weight inflation strategy, which is not currently used by state-of-the-art medical image segmentation methods, can lead to substantial improvements in performance. Moreover, we also performed systematic ablations, which further provide insights into the best strategies to adapt pre-trained Transformers from 2D to 3D through weight inflation. Our method can become a standard strategy utilized by all future works on Transformer-based models for 3D medical images to maximize performance.

Our best practice is summarized in Figure 1 and below:

- Find a vision Transformer model pre-trained on natural images via a combination of self-supervised and supervised learning.

- Adapt pre-trained weights in the embedding layer using the centering inflation strategy by transferring weights to the center-most slice and initializing all other weights to zero.

- Parse 3D input into windows of a few neighbor slices, and segment only the center slice. Aggregate 3D segmentation with predictions from all windows.

2. Related Works

Vision Transformers. Originally proposed for machine translation (Vaswani et al., 2017), Transformer based architectures have started to flourish in computer vision in recent years. ViT (Dosovitskiy et al., 2020) split images into patches and create patch embeddings as inputs to the Transformer model, which lays the foundation to use Transformer architectures for visual recognition. However, ViT requires more training data to surpass convolutional neural networks. More recent works explored different self-supervised pre-training strategies to reduce the requirement of large-scale labeled datasets. Caron et al. (Caron et al., 2021) proposed DINO, a self-distillation contrastive framework. Drawing inspiration from BERT in the NLP field, Bao et al. (Bao et al., 2021) proposed BEiT, which is pre-trained to predict masked image patches given context. These approaches achieve state-of-the-art performances in image classification and semantic segmentation.

Several works have explored leveraging Transformers for video inputs by utilizing weights from pre-trained image models (Carreira and Zisserman, 2017). ViViT (Arnab et al., 2021) and Video SwinTransformer (Liu et al., 2021b) explored inflation and centering strategies to initialize weights of video Transformers from pre-trained image Transformers. In this work, we systematically investigate the benefit of transfer learning from these Transformer models under significant domain shifts, and compare the effectiveness of transferring from different sources and pre-training objectives.

Transformers for 3D Medical Image Segmentation. Following successful adaptations of Transformers in vision, some works in medical image segmentation also achieved state-of-the-art results by utilizing Transformers. Earlier work such as Tran-

sUNet (Chen et al., 2021) or CoTr (Xie et al., 2021) still relied on convolutional layers for the Transformer model. More recent methods, such as SwinUNet (Cao et al., 2021) and UNETR (Xie et al., 2021), use a purely Transformer-based encoder. To improve the information flow of the segmentation framework, different mechanisms such as context bridge (Huang et al., 2021) or interleaved layers (Gao et al., 2021) are designed.

Based on their model input dimension, these prior works can be grouped into two categories: 3D approaches that develop 3D model architectures directly predicting 3D segmentation (Xie et al., 2021; Hatamizadeh et al., 2021); and 2D approaches that split 3D images into 2D slices and stack 2D predictions to form the 3D segmentation (Chen et al., 2021; Cao et al., 2021; Huang et al., 2021).

3D segmentation benefits from having more depth-wise contextual information, which is especially important for medical images where awareness of anatomical structures is crucial for identifying segmentation boundaries. However, due to the limited availability of pre-trained 3D Transformers, all prior 3D approaches randomly initialized their models and used no transfer learning. On the contrary, 2D approaches can easily utilize Transformer weights pre-trained on natural images, but lack the spatial context awareness of 3D models. In this work, we attempt to address the shortcomings of existing approaches by combining the benefit of transfer learning of 2D approaches and depth-wise information of 3D approaches.

3. Method

In this section, we first introduce semantic segmentation (Sec. 3.1) and vision Transformer (Sec. 3.2), and then present strategies for adapting Transformers pre-trained on 2D image to 3D medical image inputs (Sec. 3.3).

3.1. Semantic Segmentation

Semantic segmentation is the task of assigning a class label for each pixel (voxel for 3D) of an input image. Most prior works use *encoder-decoder* frameworks for semantic segmentation. The encoder progressively extracts higher-level features that capture the input image's global information. Then the decoder utilizes these features and progressively reconstructs fine-grained pixel/voxel-level predictions. Therefore, the quality of the features extracted by the encoder directly impacts the segmentation quality.

CNNs have been the dominant choices of the encoder in semantic segmentation models for years (Long et al., 2015; Ronneberger et al., 2015). Recently, several studies have reported improvements in performance for both natural and medical image segmentation by using Transformers as encoders (Liu et al., 2021a; Bao et al., 2021; Zheng et al., 2021; Chen et al., 2021; Cao et al., 2021). In this work, we compare different pre-trained Transformers as encoders and study the most effective way to perform transfer learning for 3D medical images.

3.2. Vision Transformer

Transformer was first introduced in (Vaswani et al., 2017) and adapted to visual inputs in (Dosovitskiy et al., 2020). Unlike CNNs, Transformer uses the self-attention mechanism to aggregate information from the input image, which can be viewed as more expressive convolutions and can capture long-range dependencies (Cordonnier et al., 2019). Empirical results show that vision Transformers achieve state-of-the-art performances across many vision tasks when pre-trained on large-scale natural image datasets (Dosovitskiy et al., 2020; Liu et al., 2021a; Caron et al., 2021; Bao et al., 2021).

Patch Partitioning and Embedding Layer. Vision Transformers first use a

patch partitioning and embedding layer to convert a visual input into patch embeddings to reduce input dimensionality. Specifically, for a 2D image with $H \times W$ pixels, the image is partitioned into multiple smaller patches of $P \times P$ pixels ($P \in \{16, 32\}$ in practice). Similarly, 3D images of size $H \times W \times D$ can be partitioned into patches of $P \times P \times P$ voxels. A linear layer is then applied to each patch's flattened pixel/voxel values to generate input embeddings to the Transformer encoder. By using patch partitioning, the number of inputs $T = \frac{H}{P} \times \frac{W}{P}$ ($T = \frac{H}{P} \times \frac{W}{P} \times \frac{D}{P}$ for 3D) is much smaller than the number of pixels/voxels, which makes the computation tractable for subsequent Transformer encoder layers given the $O(T^2)$ computational cost. Notably, the input patch partitioning and embedding layer is equivalent to strided convolution, where the kernel size and stride size are both equal to $P \times P$ ($P \times P \times P$ for 3D). This is important as we introduce weight inflation strategies to adapt Transformers from 2D to 3D in Sec 3.3.

Transformer Encoder. These patch embeddings $H^0 \in \mathbb{R}^{T \times D}$ are then provided as a sequence input into a *Transformer encoder* that generates contextualized representations for each patch $H^L \in \mathbb{R}^{T \times D}$. The encoder is constructed with a stack of L multi-head self-attention layers, which facilitates the aggregation of information from all the other input patches. After information exchange from every layer in a sequential manner $H^l = \text{Layer}_l(H^{l-1}) \in \mathbb{R}^{T \times D}$, the output vectors of the last layer H^L are used as contextualized representations for the input patches. We refer to (Rush, 2018; Vaswani et al., 2017) for a more detailed description of the Transformer model.

Variations of Vision Transformers. While many vision Transformers' variations exist, they all have similar architectures as described above. In this work, we compare four variations: DINO (Caron et al., 2021), BEIT (Bao et al., 2021), SwinT (Liu et al., 2021a), and VideoSwinT (Liu et al., 2021b). DINO and BEiT have the same architectures as the original vision Transformer (Dosovitskiy et al., 2020) and only differ in the self-supervised pre-training objectives. SwinT proposes shifted window attention and hierarchical feature size shrinking to introduce inductive bias of locality. VideoSwinT is an adaptation of SwinT for video inputs. As each variation has many size configurations, we use all the base models with around 85M parameters for fair comparisons.

3.3. Adapting Transformers from 2D to 3D through Weight Inflation

Many variations of Transformers that have been pre-trained on existing large-scale datasets are publicly available for transfer learning (Caron et al., 2021; Bao et al., 2021; Liu et al., 2021a,b). However, direct utilization of these Transformers is non-trivial, due to the difference in the dimensionality between 2D natural images and 3D medical images. Most existing solutions either split 3D images into 2D slices and predict each slice independently, thereby losing crucial depthwise information (Chen et al., 2021; Cao et al., 2021). Other solutions modify the Transformer architecture to support 3D inputs, but the modification makes pre-trained weights not directly applicable (Xie et al., 2021; Hatamizadeh et al., 2021).

We combine both advantages inspired by recent works in video understanding. Videos are collections of 2D image frames organized along the temporal axis, similar to medical images where 2D slices are organized along the depth axis. I3D (Carreira and Zisserman, 2017) proposed to transfer CNNs pre-trained on 2D images to 3D video inputs by inflating the convolutional weights along the temporal axis. As the input patch partitioning and

Table 1: *Details of the twelve publicly available 3D medical image segmentation datasets we used, including BCV, ACDC, and 10 MSD datasets.* These datasets cover major anatomical structures of the body and different modalities. We use the BCV dataset to develop our methods and compare them with existing methods, and use the remaining eleven datasets to test the generalizability of our best practices.

Data	Modality	Raw Reso.	Input Reso.	Clip Range	Label	Split
BCV	CT	$512 \times 512 \times 126$	$512 \times 512 \times 5$	[-175, 250]	13	18/12/0
ACDC	CT	$220 \times 247 \times 10$	$512 \times 512 \times 5$	[-175, 250]	3	70/10/20
Brain	MRI	$240 \times 240 \times 155$	$240 \times 240 \times 5$	N/A	3	387/72/25
Heart	MRI	$320 \times 320 \times 114$	$320 \times 320 \times 5$	N/A	1	16/3/1
Liver	CT	$512 \times 512 \times 447$	$512 \times 512 \times 5$	[-175, 250]	2	104/20/7
Hippocampus	MRI	$34 \times 49 \times 36$	$240 \times 240 \times 5$	N/A	2	208/39/13
Prostate	MRI	$320 \times 320 \times 19$	$320 \times 320 \times 5$	N/A	2	25/5/2
Lung	CT	$512 \times 512 \times 282$	$512 \times 512 \times 5$	[-175, 250]	1	50/9/4
Pancreas	CT	$512 \times 512 \times 94$	$512 \times 512 \times 5$	[-175, 250]	2	224/42/15
Vessel	CT	$512 \times 512 \times 69$	$512 \times 512 \times 5$	[-175, 250]	2	242/45/16
Spleen	CT	$512 \times 512 \times 88$	$512 \times 512 \times 5$	[-175, 250]	1	32/6/3
Colon	CT	$512 \times 512 \times 110$	$512 \times 512 \times 5$	[-175, 250]	1	100/19/7

embedding layer in Transformer is equivalent to CNN, we adopt the weight inflation strategy to initialize the 3D Transformers.

We compare two inflation strategies: *average inflation* and *centering inflation*. For average inflation, we copy the weights of the input layer K times in the depth axis and divide them by K. This setting assumes input slices are similar within a certain range of depths, and the model treats all the input slices equally at the beginning. For centering inflation, we transfer pre-trained weights for the center-most slice and initialize all other weights to zero. For this setting, the model uses only information from the center slice at the beginning and progressively learns to contextualize information from all the neighbor slices. Both strategies keep the mean and variance of the input to Transformer encoder, allowing more effective transfer learning.

Another difference is in the input channel. Transformers pre-trained on colored natural images have three channels. Similarly, We reduce the channel to one by modifying the in-put layer weight, where we take the sum over the input channel axis. We further use average inflation in the input channel for medical images with more than one input channel such as MRIs. These modifications still keep the mean and variance of the input unchanged for the Transformer encoder. After adapting the weight in the input patch partition and embedding layer, the weights of Transformer encoder layers can be directly initialized from 2D pre-trained Transformers since input shapes are aligned.

4. Results

4.1. Dataset

We use 12 publicly available 3D medical image datasets in this study: Beyond the Cranial Vault (BCV) multi-organ segmentation dataset (BCV), Automated Cardiac Diagnosis Challenge (ACDC) (ACD), and 10 datasets from Medical Segmentation Decathlon (MSD) (Antonelli et al., 2021). These datasets cover major anatom-

Table 2: *Segmentation results on the BCV dataset. (a)* Our approach achieves state-of-the-art performances by combining the advantages of transfer learning (T) and depth information (D). DSC is the Dice coefficient averaged on the 8 major organs. *(b)* Transfer effectiveness from models pre-trained with different sources and objectives. NI, NV, and MI are natural images, natural videos, and medical images. SL and SSL are supervised learning and self-supervised learning.

(a) **Main Result**				(b) **Transfer Source and Objective**			
Method	**T**	**D**	**DSC**	**Encoder**	**Source**	**Objective**	**DSC**
TransUNet (2021)	✔	✘	84.36	BEiT-Random	-	-	74.70
SwinUNet↓ (2021)	✔	✘	81.12	CT-DINO	MI (RSNACT)	SSL	82.56
UNETR (2021)	✘	✔	80.07	BEiT-SSL (2021)	NI (IN22K)	SSL	84.41
				DINO (2021)	NI (IN1K)	SSL	84.60
Ours w/o T&D	✘	✘	74.00	SwinT (2021a)	NI (IN22K)	SL	85.57
Ours w/o T	✘	✔	74.70	VideoSwinT (2021b)	NV (K600)	SL	85.94
Ours w/o D	✔	✘	85.18	BEiT (2021)	NI (IN22K)	SSL+SL	**87.13**
Ours	✔	✔	**87.13**				

ical structures of the body and different modalities such as CT and MRI. We use the same BCV training and validation set as existing works (Chen et al., 2021; Cao et al., 2021) to develop our method and fairly compare it with other methods. We use the remaining 11 datasets to test the generalizability of our best practice. Details of the datasets such as labels, modality, resolution, data split are available in Table 1.

4.2. Main Result

Most existing works such as TransUNet (Chen et al., 2021) and SwinUNet (Cao et al., 2021) that use Transformers for 3D medical image segmentation are slice-based approaches, which split a 3D image into 2D slices and predict segmentation independently on each slice. This approach enables the direct utilization of pre-trained weights for transfer learning, but none of these works study the effect of transfer learning by comparing their methods with randomly initialized encoders. We find that *transfer learning significantly boosts the*

segmentation performance even under significant domain shifts. Using the pre-trained encoder leads to 11.18% improvements in DSC as compared to the same randomly initialized encoder (Table 2(a))[2].

Recent works such as CoTr (Xie et al., 2021) or UNETR (Hatamizadeh et al., 2021) modify the Transformer architecture to support 3D inputs to leverage depth information. However, they all use randomly initialized 3D Transformers as encoders, which sacrifices the clear benefit of transfer learning as discussed above. Due to the different amount of training data and number of labels reported in their papers[3], we reproduced UNETR using its officially released codes for a fair comparison. From Table 2, we can see UNETR underperforms TransUNet or Swi-

---

2. SwinUNet↓ indicates an input resolution of 384, while other models use an input resolution of 512. TransUNet and SwinUNet results are from their original papers, while other results are from our experiments.

3. CoTr use 21 training data from the BCV dataset and reported DSC based on 11 organs, while UNETR uses 28 + external training data and the result is based on an ensemble of four models.

Table 3: *Fine-grained segmentation results of different organs on the BCV dataset.* We report performances on Aorta, Gallbladder, Kidney (Left), Kidney (Right), Liver, Pancreas, Spleen, and Stomach, respectively. The most significant improvements of our method are Gallbladder, Pancreas, and Stomach segmentation, which can be interpreted as the larger the variation between neighbor slices, the more improvement from incorporating depth information through weight inflation.

Method	Aor	Gal	KidL	KidR	Liv	Pan	Spl	Sto
TransUNett (2021)	90.68	71.99	86.04	83.71	95.54	73.96	88.80	84.20
SwinUNet$_\downarrow$ (2021)	87.07	70.53	84.64	82.87	94.72	63.73	90.14	75.29
UNETR (2021)	86.30	63.79	84.94	83.93	95.94	58.01	88.95	78.76
Ours w/o T&D	82.71	61.34	80.73	76.25	94.37	46.52	82.62	67.46
Ours w/o T	81.26	63.78	80.90	75.41	94.82	53.67	82.15	65.64
Ours w/o D	90.22	67.51	93.84	91.40	96.27	66.62	94.22	81.33
Ours	90.29	75.20	93.70	91.40	96.17	71.82	93.81	84.62
# Voxels	441936	118445	887310	879180	9762041	504539	1386739	2641170
# Slices	827	131	362	350	631	316	332	421
Slice Variation	90.33	69.84	82.80	82.57	88.93	75.91	83.37	84.94

nUNet, but it achieves 6.07% improvements in DSC compared to the randomly initialized slice-based model, demonstrating *the importance of depth information.*

To combine the advantages of *transfer learning* and *depth information*, we use the weight inflation strategy introduced in Sec. 3.3 to initialize 3D Transformers. We observe consistent improvements with this initialization over random, but we find that the weight inflation depth should not be too large (Sec. 4.4). Our best model uses a window-based approach. Specifically, we split 3D images into small windows along the depth axis, where each window consists of a few (e.g., 5) neighbor slices. We aggregate all the window predictions to form the final prediction. *By incorporating shallow depth information, our method achieves 1.95% improvements in DSC compared to slice-based methods.* More importantly, *such improvement is at only 0.75% increased computational cost* (213.3 GFLOPS before adaptation vs. 215.0 after).

We also evaluate our method stability by running our model and baselines on the BCV dataset three times with three random seeds (1234, 5678, 910). Our model (i.e., Ours in Table 2(a)) achieves 86.65 ± 0.35 (87.13, 86.52, 86.29 for three runs) DSC, while the model without weight inflation (i.e., Ours w/o D in Table 2(a)) only achieves 85.14 ± 0.20 (85.18, 84.88, 85.36 for three runs) DSC. The results indicate *a 98% level of significance* using the paired t-test.

To better evaluate the improvements of our method, we extend Table 2 and show the DSC results for each of the 8 organs in Table 3. We find that the most significant improvements from weight inflation are Gallbladder, Pancreas, and Stomach segmentation. We compute different statistics of each organ to understand the reason, including the number of voxels, the number of occurring slices, and the variation between slices (measured by the average DSC between labels of neighbor slices). We find that *the larger the variation between neighbor slices,*
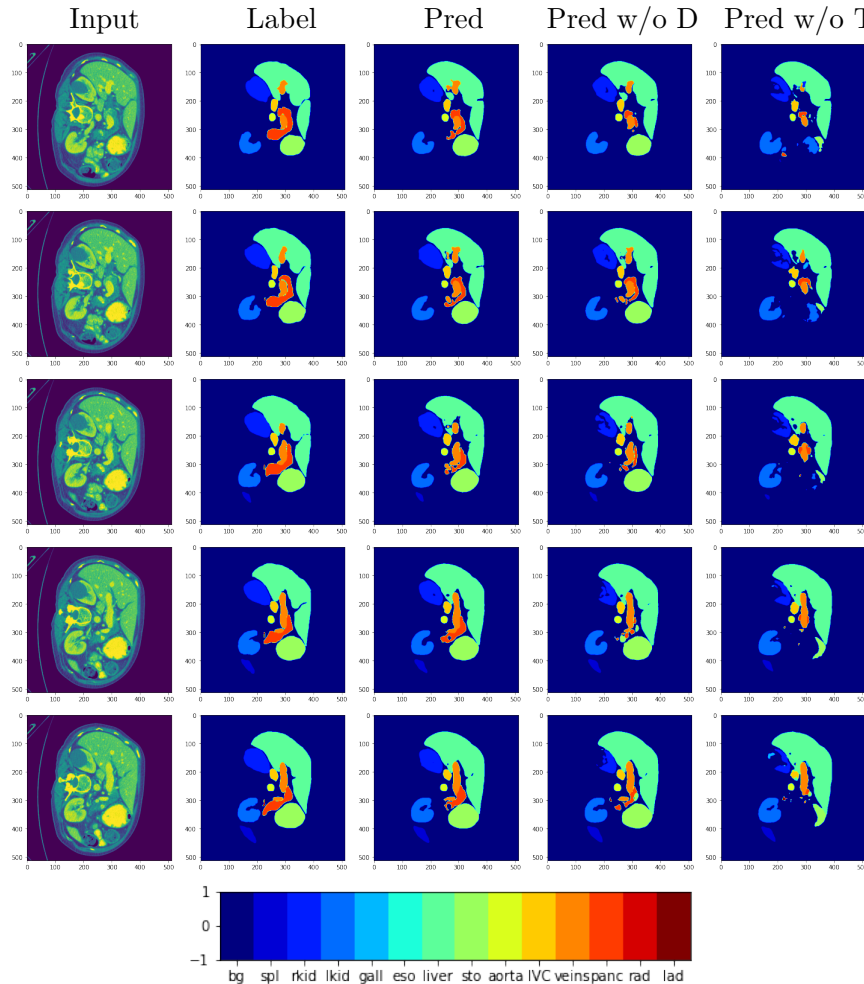
398

Figure 2: *Transfer learning reduces noise in local predictions, and depth information further improves prediction consistency across slices.* We show five consecutive input images, ground-truth labels, predictions from *Ours, Ours w/o D, Ours w/o T&D*.

the more improvements from incorporating depth information through weight inflation. This can be interpreted as *inter-slice prediction consistencies are significantly improved with depth information* based on the qualitative result shown in Figure 2.

4.3. Pre-training Source and Objective

Since pre-training sources may significantly impact the transfer performance, we com-

pare the transfer effectiveness of various vision Transformers pre-trained with different sources (natural images, medical images, natural videos) and objectives (supervised learning, self-supervised learning, combination of both) (Table 2(b)). To fairly compare models, we carefully select same-size models with similar computations. SwinT (Liu et al., 2021a) is pre-trained on natural image dataset ImageNet-22K (Deng et al., 2009) via supervised learning. DINO (Caron

Table 4: *Best practice of weight inflation on the BCV dataset.* Initializing the model with the centering inflation strategy and only predicting the center slice's segmentation from a small number of neighbor slices leads to the best performance.

(*a*) **Prediction Target**				(*b*) **Weight Initialization**			
Prediction	All	**Center**		Initialization	Random	Average	**Centering**
DSC	73.93	**76.33**		DSC	72.49	76.33	**84.30**
(*c*) **Input Stride**				(*d*) **Input Slices**			
Input Stride	**1**	2	3	Input Slices	1	**5**	11	21
DSC	**84.74**	84.30	77.68	DSC	83.34	**84.74**	83.83	77.52

et al., 2021) is pre-trained on ImageNet-1K via self-supervised contrastive learning. BEiT-SSL (Bao et al., 2021) is pre-trained on ImageNet-22K via self-supervised masked image modeling, and BEiT is further fine-tuned on the same dataset via supervised learning. VideoSwinT (Liu et al., 2021b) is pre-trained on natural video dataset Kinetics-600 (Carreira et al., 2018). We also pre-trained a CT-DINO on the medical image dataset RSNA-CT (Colak et al., 2021) using the same model and objective as DINO.

Interestingly, we find that *models pre-trained on medical images performs worse than models pre-trained on natural images if both models are pre-trained using self-supervised learning.* This might be attributed to the high inter-class similarity for medical images due to the nature of human anatomy, causing limitations in learning meaningful representations using self-supervised learning. Furthermore, medical image datasets are much smaller than natural image datasets. Designing effective self-supervised learning methods in medical images will be meaningful future work. Moreover, we find that *additional supervision from image labels or video labels further improved the transfer performance*, as models pre-trained via supervised learning generally achieved better performances than models pre-trained via self-supervised learning.

Since our goal is to explore the best strategy to leverage existing large-scale pre-trained 2D vision Transformers given their strong adaptability to various tasks and public availability, we did not perform additional pre-training. Thoroughly exploring pre-training on different sources with different objectives and data amounts is a meaningful future direction but requires extensive computational resources and costs.

4.4. Best Practice of Weight Inflation

We conduct ablations on different inflation settings, including prediction target, weight initialization, stride, and the number of input slices (Table 4)[4]. We find that *initializing the model with the centering inflation strategy and only predicting the center slice's segmentation leads to the best performance.* This setting is similar to residual connection (He et al., 2016), where the model gradually learns to utilize additional depth information, and it should not be worse than the model without depth information. Moreover, we find that *including a few neighbor slices is enough,* because neighbor slices are more similar to the center slice than non-

---

4. Here DSC is averaged on all the 14 labels instead of 8 major organs in Table 2.

Table 5: *Generalizability of best practice.* To verify the generalization of our best practice that adapts Transformers from 2D to 3D, we compare our adapted Transformers with original Transformers on 11 additional datasets and observe improvements over 10 datasets. DSC excluding background on the test set is reported. We also include other baseline performances for reference, but numbers are not fully comparable due to no standardized data split available for these datasets (+ indicates same data split ratio but not same split; * indicates different data split ratios).

Method	ACDC	Bra	Hea	Liv	Hip	Pro	Lun	Pan	Ves	Spl	Col
SwinUNet[+] (2021)	90.00	-	-	-	-	-	-	-	-	-	-
UNETR[+] (2021)	-	71.1	-	-	-	-	-	-	-	96.4	-
nnUNet* (2021)	91.79	74.11	93.28	79.71	88.91	75.37	72.11	67.45	68.37	96.38	45.53
Ours w/o D	92.62	79.30	91.22	87.24	86.24	74.29	62.01	58.60	67.95	95.12	46.28
Ours	92.69	80.13	90.78	87.67	86.83	75.20	70.20	59.94	70.42	95.56	51.70

neighbor slices are and provide more information for the model to predict accurate segmentation. Including too many slices may lead to overfitting. This finding also applies to VideoSwinT, where we observe 81.10 and 81.83 DSC for 9 and 5 slices, respectively.

4.5. Generalization of Best Practice

To verify the generalizability of our best practice — adapting pre-trained vision Transformers from 2D to 3D using the strategies in Table 4, we train and evaluate our method on 11 additional 3D medical image datasets of different anatomical regions and imaging modalities. From Table 5, we observe that *weight-inflated vision Transformers (Ours) achieve consistently better performances on 10 of 11 datasets compared to the original Transformers (Ours w/o D), which clearly demonstrates the effectiveness of our proposed method.* Also, these results are obtained with just a single set of hyperparameters tuned on the BCV dataset, *showing the robustness of our method.*

Since our model uses all the hyperparameters tuned on the BCV dataset, continuing tuning hyperparameters on each dataset should lead to further improvements. For example, since the hyperparameters of our method are tuned on the CT dataset, the optimal hyperparameters may be very different for MRI datasets, thus MRI datasets such as Bra, Hea, Hip, and Pro show smaller gains than other CT datasets. However, since the main purpose of our paper is to show the importance of weight inflation, not to achieve different state-of-the-arts, we leave the relation between data characteristics and optimal hyperparameters to future work.

5. Conclusion

In this work, we investigated adapting pre-trained Transformers for 3D medical image segmentation via simple yet effective weight inflation strategies. Our approach achieved consistent improvements on 12 datasets with only 0.75% increased computational cost, which can become a standard strategy easily utilized by all work on Transformer-based models for 3D medical images, to maximize performance.

Acknowledgments

References

Automated cardiac diagnosis challenge (acdc). https://www.creatis.insa-lyon.fr/Challenge/acdc/.

Beyond the cranial vault (bcv) multi-organ segmentation. https://www.synapse.org/#!Synapse:syn3193805/wiki/217789.

José Luis Alonso-Martínez, FJ Anniccherico Sánchez, and MA Urbieta Echezarreta. Delay and misdiagnosis in sub-massive and non-massive acute pulmonary embolism. *European Journal of Internal Medicine*, 2010.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

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. *arXiv preprint arXiv:2105.05537*, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

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.

Errol Colak, Felipe C Kitamura, Stephen B Hobbs, Carol C Wu, Matthew P Lungren, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Robyn L Ball, George Shih, Anouk Stein, et al. The rsna pulmonary embolism ct dataset. *Radiology: Artificial Intelligence*, 2021.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *ICLR*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Andre Esteva, Katherine Chou, Serena Ye-ung, Nikhil Naik, Ali Madani, Ali Mot-taghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 2021.

Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer ar-chitecture for medical image segmentation. In *MICCAI*, 2021.

Ali Hatamizadeh, Dong Yang, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. *arXiv preprint arXiv:2103.10504*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.

William R Hendee, Gary J Becker, James P Borgstede, Jennifer Bosma, William J Casarella, Beth A Erickson, C Douglas Maynard, James H Thrall, and Paul E Wallner. Addressing overutilization in medical imaging. *Radiology*, 2010.

Janneke MT Hendriksen, Marleen Koster-van Ree, Marcus J Morgenstern, Ruud Oudega, Roger EG Schutgens, Karel GM Moons, and Geert-Jan Geersing. Clinical characteristics associated with diagnostic delay of pulmonary embolism in primary care: a retrospective observational study. *BMJ Open*, 2017.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learn-ing: a systematic review and implemen-tation guidelines. *NPJ digital medicine*, 2020.

Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An ef-fective medical image segmentation trans-

former. *arXiv preprint arXiv:2109.07162*, 2021.

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*, 2021.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yix-uan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierar-chical vision transformer using shifted win-dows. *arXiv preprint arXiv:2103.14030*, 2021a.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021b.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional net-works for biomedical image segmentation. In *MICCAI*, 2015.

Alexander M Rush. The annotated trans-former. In *ACL NLP-OSS Workshop*, 2018.

Rebecca Smith-Bindman, Diana L Miglioretti, and Eric B Larson. Ris-ing use of diagnostic medical imaging in a large integrated health system. *Health affairs*, 2008.

Ashish Vaswani, Noam Shazeer, Niki Par-mar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridg-

ing cnn and transformer for 3d medical image segmentation. *arXiv preprint arXiv:2103.03024*, 2021.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.