AdvMIM: Adversarial Masked Image Modeling for Semi-Supervised Medical Image Segmentation

Lei Zhu[1✉], Jun Zhou[1], Rick Siow Mong Goh[1], and Yong Liu[1]

Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore
zhu_lei@ihpc.a-star.edu.sg

Abstract. Vision Transformer (ViT) has recently gained tremendous popularity in medical image segmentation task due to its superior capability in capturing long-range dependencies. However, transformer requires a large amount of labeled data to be effective, which hinders its applicability in annotation scarce semi-supervised learning scenario where only limited labeled data is available. State-of-the-art semi-supervised learning methods propose combinatorial CNN-Transformer learning to cross teach a transformer with a convolutional neural network (CNN), which achieves promising results. However, it remains a challenging task to effectively train the transformer with limited labeled data. In this paper, we propose an adversarial masked image modeling (**AdvMIM**) method to fully unleash the potential of transformer for semi-supervised medical image segmentation. The key challenge in semi-supervised learning with transformer lies in the lack of sufficient supervision signal. To this end, we propose to construct an auxiliary masked domain from original domain with masked image modeling and train the transformer to predict the entire segmentation mask with masked inputs to increase supervision signal. We leverage the original labels from labeled data and pseudo-labels from unlabeled data to learn the masked domain. To further benefit the original domain from masked domain, we provide a theoretical analysis of our method from a multi-domain learning perspective and devise a novel adversarial training loss to reduce the domain gap between the original and masked domain, which boosts semi-supervised learning performance. We also extend adversarial masked image modeling to CNN network. Extensive experiments on three public medical image segmentation datasets demonstrate the effectiveness of our method, where our method outperforms existing methods significantly. Our code is publicly available at https://github.com/zlheui/AdvMIM.

Keywords: Adversarial Training · Masked Image Modeling · Semi-Supervised Segmentation.

1 Introduction

Medical image segmentation is an important task for computer assisted diagnosis, treatment planning, and intervention. With the recent advancement of vision transformer [9] and its exceptional capability in capturing long-range dependencies, there is growing interest in the medical domain to leverage transformer for medical image segmentation task [6, 5]. However, vision transformer is even more annotation hungry than convolutional neural network (CNN) [22]. Semi-supervised learning methods aim to leverage a large amount of unlabeled data together with a limited amount of labeled data for learning a segmentation network to reduce the annotation cost. Existing semi-supervised learning methods can be categorized into pseudo-labeling based methods [1, 24, 7], consistency regularization based methods [21, 27, 18, 16, 23, 2], deep co-training based methods [19, 7, 29], and adversarial training based methods [28, 11]. However, benchmarking results [15, 12] indicate that directly integrating these methods with transformer leads to poor performance, likely due to the annotation dependency of transformer. State-of-the-art semi-supervised learning method [15] proposes cross-teaching to leverage the complementary architectural advantages of both CNN with efficient local features learning and transformer with better long-range dependencies capturing for the task. Most recently, Huang et al. [12] propose a combinatorial CNN-Transformer learning framework at manifold space with intra-student consistency regularization and inter-student knowledge transfer, which achieves state-of-the-art performance on multiple datasets. While these methods have achieved promising results, it remains a challenging task to effectively train the transformer with limited labeled data.

In this paper, we propose an adversarial masked image modeling (**AdvMIM**) method to fully unleash the potential of transformer for semi-supervised medical image segmentation. The key challenge in semi-supervised learning with transformer lies in the lack of sufficient supervision signal. While combinatorial CNN-Transformer learning based methods [15, 12] leverage a CNN network to assign pseudo-labels to unlabeled data for training the transformer, which boosts model performance, we believe even more supervision signal is needed to effectively train the transformer. Therefore, we propose to construct an auxiliary masked domain from original domain with masked image modeling [26] and perform masked domain learning with transformer to increase the supervision signal. Specifically, masked image modeling [26] is an effective self-supervised learning method for vision transformer, where the task is to reconstruct the masked image patches. We employ the same masking operation to construct a masked domain and train the vision transformer to predict the entire segmentation mask from the masked inputs. We utilize the original labels from labeled data and pseudo-labels from unlabeled data to learn the masked domain. With the new input and new task, the transformer gains extra supervision signal for learning. To further benefit the original domain from masked domain, we provide a theoretical analysis of our framework from a multi-domain learning perspective and devise a novel adversarial training loss to reduce the domain gap between the original and masked domain, where we employ a domain discriminator to distinguish the prediction masks of both original labeled data and masked unlabeled data and adversarially train the transformer to produce more accurate

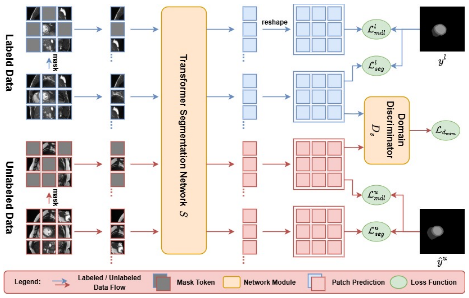

Fig. 1. Architecture and dataflow of our proposed adversarial masked image modeling method. Our method constructs an auxiliary masked domain from original domain with masked image modeling. We utilize original labels from labeled data and pseudo-labels from unlabeled data to learn the masked domain. We further propose a novel adversarial training loss to reduce the domain gap between the original and masked domain to boost semi-supervised learning performance. Note pseudo-label $\hat{y}$ is obtained from a cross-teaching CNN.

prediction masks for the unlabeled masked data so that the discriminator cannot distinguish them from those of labeled data.

In summary, we have made the following contributions in this paper: **(1).** We propose an adversarial masked image modeling method to fully unleash the potential of transformer for semi-supervised medical image segmentation; **(2).** We provide a theoretical analysis of our method from a multi-domain learning perspective and propose a novel adversarial training loss to reduce the domain gap between masked and original domain; **(3).** We extend adversarial masked image modeling to CNN network; **(4).** We perform extensive experiments to evaluate our method on three public medical image segmentation datasets, where our method outperforms existing methods significantly.

2 Methodology

In semi-supervised medical image segmentation, we are given N^l labeled data $\mathbb{D}^l = \{(x^l_i, y^l_i)\}^{N^l}_{i=1}$ and N^u unlabeled data $\mathbb{D}^u = \{x^u_i\}^{N^u}_{i=1}$, where y^l_i is the corresponding segmentation mask for x^l_i with M classes and $N^l << N^u$. Both the labeled data and unlabeled data are sampled from probability distribution P, where unlabeled data are sampled without labels. The goal is to learn an accurate segmentation model with both labeled and unlabeled data. Fig. 1 presents an overview of our proposed adversarial masked image modeling method.

2.1 Segmentation

As illustrated in Fig. 1, we train the transformer segmentation network $S : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W\times M}$ with both labeled data and pseudo-labeled unlabeled data. We follow existing methods [15, 12] to cross teach the transformer network concurrently with a convolutional neural network $C : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H \times W \times M}$, where the pseudo-label for unlabeled data is obtained through $\hat{y}_i^u = argmax(C(x_i^u))$. As pseudo-label may contain noise, we weight the loss function with its maximum predicted probability as a certainty measure [14], where low-quality pseudo-labels will have small weights especially in early training iterations. The weight is calculated as $w_i^u = max(C(x_i^u))$. The segmentation loss functions on labeled and unlabeled data for the transformer are defined as follows:

$$\mathcal{L}_{seg}^{l}(S) = \mathbb{E}_{x^l \sim \mathbb{D}^l}[H(y^l, S(x^l)) + Dice(y^l, S(x^l))], \tag{1}$$

$$\mathcal{L}_{seg}^{u}(S) = \mathbb{E}_{x^u \sim \mathbb{D}^u}[w^u H(\hat{y}^u, S(x^u)) + Dice(\hat{y}^u, S(x^u))], \tag{2}$$

where $H(\cdot)$ calculates the pixel-wise cross-entropy loss and $Dice(\cdot)$ calculates the *Dice* loss. Note the CNN is trained together with the transformer using the same loss functions except that pseudo-labels for CNN are assigned by transformer.

2.2 Masked Domain Learning

The key challenge in semi-supervised medical image segmentation with vision transformer lies in the lack of sufficient supervision signal. We propose to construct an auxiliary masked domain from the original domain with masked image modeling [26]. But instead of reconstructing the masked image patches, we propose to train the transformer to predict the entire segmentation mask with masked input, where the transformer learns to infer the semantics on masked image patches based on the visible ones. We employ the same masking operation in masked image modeling to construct masked images. We replace the masked image patches with shared learnable mask tokens with positional embedding and utilize original labels from labeled data and pseudo-labels from unlabeled data to learn the masked domain. With the new input and new task, the transformer gains extra supervision signals for learning. The masked domain learning loss for labeled and unlabeled data are defined as follows:

$$\mathcal{L}_{mdl}^{l}(S) = \mathbb{E}_{x^l \sim \mathbb{D}^l}[H(y^l, S(x^{ml}))) + Dice(y^l, S(x^{ml})], \tag{3}$$

$$\mathcal{L}_{mdl}^{u}(S) = \mathbb{E}_{x^u \sim \mathbb{D}^u}[w^u H(\hat{y}^u, S(x^{mu}) + Dice(\hat{y}^u, S(x^{mu})], \tag{4}$$

where $x^{ml} = [\mathcal{M}(x^l, \rho); \boldsymbol{T}]$ is the masked labeled data, $\mathcal{M}(\cdot, \rho)$ denotes the masking operation with mask ratio ρ, $\boldsymbol{T}$ denotes the set of mask tokens to replace the masked image patches with positional embedding, the operation $[\cdot; \cdot]$ concatenates two input vectors into a single vector, and $x^{mu} = [\mathcal{M}(x^u, \rho); \boldsymbol{T}]$. Following [26], we set $\rho = 0.7$ by default.

2.3 Adversarial Masked Domain Adaptation

We treat both the pseudo-labeled original domain and the pseudo-labeled masked domain as noisily labeled domain. Inspired by [3], we analyze our framework from a multi-domain learning perspective and present the following theorem:

Theorem 1 (Masked Domain Adaptation Theorem). *Following the problem definition, let h be a hypothesis in class $\mathcal{H}$, let γ be the pseudo-label noise ratio, denote the probability distribution of pseudo-labeled original domain as P' and the probability distribution of pseudo-labeled masked domain as Q', then for any $\delta \in (0,1)$, with probability at least $1-\delta$, for every $h \in \mathcal{H}$, we have:*

$$\epsilon_P(h) \leq \frac{1}{2}\epsilon_{P'}(h) + \frac{1}{2}\epsilon_{Q'}(h) + \frac{1}{4}d_{\mathcal{H}\varDelta\mathcal{H}}(P, Q') + \frac{1}{2}\lambda + \gamma, \tag{5}$$

where $\epsilon_P(\cdot)$ (resp. $\epsilon_{P'}(\cdot)$, $\epsilon_{Q'}(\cdot)$) measures the expectation error of a hypothesis on original (resp. pseudo-labeled original, pseudo-labeled masked) data distribution, $d_{\mathcal{H}\varDelta\mathcal{H}}(\cdot,\cdot)$ measures the distribution discrepancy between two data distributions, and $\lambda = \min_{h\in\mathcal{H}} \epsilon_P(h) + \epsilon_Q(h)$.

Proof (Sketch). We leverage *Lemma* 4 in [3] to bound the expectation error on original domain with multi-domain learning loss on both original and pseudo-labeled masked domain. We apply triangle inequality to bound both the expectation error of original domain and the optimal error between original and pseudo-labeled masked domain with pseudo-label noise ratio. After term combination and rearrangement, we derive the final bound.

The theorem upper bounds expectation error on original domain with (1). expectation error on pseudo-labeled original domain; (2). expectation error on pseudo-labeled masked domain; (3). the domain gap between original and masked domain; (4). pseudo-label noisy ratio; and (5). the non-optimizable optimal error between original and pseudo-labeled masked domain that is assumed to be small [3]. Our segmentation loss and masked domain learning loss functions minimize term (1) and term (2) respectively. We weight the loss functions with certainty measures to minimize term (4). Except for the non-optimizable term (5), our theorem further indicates that it is necessary to minimize term (3), the domain gap between original and masked domain to fully bound the expectation error on original domain.

To this end, we introduce a domain discriminator $D_s : \mathbb{R}^{H\times W\times M} \rightarrow \mathbb{R}$, which takes the prediction masks from both original labeled data and masked unlabeled data as input and outputs the domain prediction. We adversarially train the transformer to confuse the domain discriminator, where the transformer and the domain discriminator play a two-player min-max game following the GAN [10] framework. At the optimal, the transformer aligns the masked unlabeled data distribution towards the original labeled data distribution to reduce the domain gap between masked and original domain so that the discriminator cannot distinguish the prediction masks between them anymore. Following [17], we adopt the least squares loss for GAN training to enhance the training tability. The adversarial masked image modeling loss is defined as follow:

$$\begin{aligned}\mathcal{L}_{d_{mim}}(D_s) = \mathbb{E}_{x^l\sim\mathbb{D}^l}[(D_s(S(x^l)) - 1)^2] \\ + \mathbb{E}_{x^u\sim\mathbb{D}^u}[(D_s(S(x^{mu})))^2],\end{aligned} \tag{6}$$

$$\mathcal{L}_{adv_{mim}}(S) = \mathbb{E}_{x^u\sim\mathbb{D}^u}[(D_s(S(x^{mu})) - 1)^2]. \tag{7}$$

Discussion. Note different from existing adversarial training based semi-supervised learning method [28], which performs adversarial training to reduce the domain gap between labeled and unlabeled data, our method constructs and learns an auxiliary masked domain and performs adversarial training to reduce the domain gap between original and masked domain on labeled and masked unlabeled data.

2.4 Extension to CNN

We propose to extend adversarial masked image modeling to CNN network to improve its learning with limited labeled data, which in turn can benefit the learning of transformer through the cross-teaching process. Specifically, we apply the same masking operation and perform masked domain learning and adversarial masked domain adaptation on CNN network. We add a domain discriminator D_c to reduce the domain gap between original and masked domain in CNN branch. Due to the symmetry of transformer and CNN branch, the same loss functions are defined for CNN network. Without loss of generality, the overall objective of our framework is defined as follows:

$$\begin{aligned} &\min_{S,C} \mathcal{L}^l_{seg} + \mathcal{L}^u_{seg} + \mathcal{L}^l_{mdl} + \mathcal{L}^u_{mdl} + \lambda_{adv}\mathcal{L}_{adv_{mim}}, \\ &\min_{D_s,D_c} \mathcal{L}_{d_{mim}}, \end{aligned} \tag{8}$$

where λ_{adv} is balancing weight which is empirically set to be 0.001.

3 Experimental Analysis

Datasets. We evaluate the effectiveness of our framework on three public datasets, namely Automated Cardiac Diagnosis Challenge (ACDC) [4], Synapse [13], and International Skin Imaging Collaboration (ISIC) [8]. ACDC contains 100 magnetic resonance imaging (MRI) scans of three organs. Following [15], we adopt 70, 10 and 20 cases for training, validation and testing. We evaluate with 3% and 10% partitions of training data as labeled data, while the rest training data as unlabeled data for semi-supervised segmentation. Synapse consists of 30 computed tomography (CT) scans annotated with eight abdominal organs. Following [12], we adopt 18 and 12 cases for training and testing and evaluate with 15% and 30% partitions of the training data. ISIC is a skin lesion segmentation dataset including 2,594 dermoscopy images, with 1,838 training images and 756 validation images. We experiment with 3% and 10% partitions of training data.
Implementation. In all experiments, we adopt Swin-UNet [5] as the transformer segmentation network and UNet [20] as the convolutional neural network. The UNet is only used for complementary training and not used for final prediction. We implement the domain discriminator with a five-layer CNN network. We train our framework with SGD optimizer for 30,000 iterations, where the initial learning rate is 0.05, momentum is 0.9 and weight decay is 1e-4. The batch size is 16 with half labeled and half unlabeled images. Following [12], we randomly crop

Table 1. Ablation study on ACDC (10%) and Synapse (15%) in Dice (%) and HD (mm). The best performance is marked in **bold**.

Method	$\mathcal{L}^l_{seg}$		$\mathcal{L}^u_{seg}$		$\mathcal{L}^l_{mdl}$		$\mathcal{L}^u_{mdl}$		$\mathcal{L}_{adv_{mim}}$		ACDC (10%)		Synapse (15%)	
	S	C	S	C	S	C	S	C	S	C	Dice	HD	Dice	HD
Labeled only	✓										79.6	4.1	45.7	43.1
+Cross Teaching$_{(\text{weighted})}$	✓	✓	✓	✓							86.6	2.5	63.1	26.7
+Masked Domain Learning	✓	✓	✓	✓	✓		✓				88.2	1.8	65.0	23.5
+Adversarial Masked Domain Adaptation	✓	✓	✓	✓	✓		✓		✓		88.9	**1.3**	65.8	22.9
AdvMIM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**89.0**	**1.3**	**66.3**	**22.7**

Table 2. Comparison with SoTA methods on ACDC, Synapse, and ISIC in Dice (%) and HD (mm). The best results are in **bold**, and the second-best results are underlined.

Method	ACDC (3%)		ACDC (10%)		Synapse (15%)		Synapse (30%)		ISIC (3%)		ISIC (10%)	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD	Dice	HD	Dice	HD
MT [21]	56.6	34.5	81.0	14.4	49.7	69.4	61.1	63.8	72.8	37.4	73.4	34.0
UA-MT [27]	61.0	25.8	81.5	14.4	51.3	93.4	57.8	63.9	73.0	38.6	73.4	33.2
EM [24]	60.2	24.1	79.1	14.5	49.5	72.7	59.7	63.8	72.3	36.3	72.7	39.3
DCT [19]	58.2	26.4	80.4	13.8	51.0	77.0	60.6	64.2	72.9	40.6	76.0	35.7
CCT [18]	58.6	27.9	81.6	13.1	40.2	75.9	57.6	69.9	67.7	42.2	72.3	31.7
CPS [7]	60.3	25.5	83.3	11.0	47.9	66.2	60.7	69.0	68.6	44.4	74.3	35.7
ICT [23]	58.1	22.8	81.1	11.4	52.7	70.5	62.7	59.6	73.2	37.2	75.3	34.6
DAN [28]	52.8	32.6	79.5	14.6	47.0	93.3	58.3	73.3	69.5	39.5	72.4	30.4
URPC [16]	56.7	31.4	82.9	10.6	48.9	69.6	59.7	66.0	70.3	39.3	75.8	32.8
CTCT [15]	70.4	12.4	86.4	8.6	60.4	45.4	68.7	44.3	71.3	43.2	76.0	37.3
SSNet [25]	70.5	17.4	85.3	10.6	58.1	47.3	66.8	34.9	72.8	40.8	75.8	32.8
ICT-Med [2]	56.3	22.6	83.7	13.1	51.5	62.0	61.2	59.1	71.4	39.2	74.9	33.1
M-CnT [12]	75.3	10.7	88.4	4.4	65.3	32.6	71.4	31.2	77.9	32.1	81.1	24.4
AdvMIM	**85.4**$_{(10.1\uparrow)}$	**2.0**$_{(8.7\downarrow)}$	**89.0**$_{(0.6\uparrow)}$	**1.3**$_{(3.1\downarrow)}$	**66.3**$_{(1.0\uparrow)}$	**22.7**$_{(9.9\downarrow)}$	**74.8**$_{(3.4\uparrow)}$	**15.5**$_{(15.7\downarrow)}$	**79.8**$_{(1.9\uparrow)}$	**22.2**$_{(9.9\downarrow)}$	**81.8**$_{(0.7\uparrow)}$	**19.5**$_{(4.9\downarrow)}$

a patch with size of 224×224 as the input. We perform standard data augmentation to avoid overfitting, including random flip and rotation. We employ two commonly-used metrics, the Dice coefficient (Dice) and the Hausdorff Distance (HD) to quantitatively evaluate the segmentation performance.

Ablation Study. In Table 1, we present the ablation study of different components of our method. As can be observed, the baseline labeled only method performs poorly. Our weighted cross-teaching loss, which integrates certainty measures to cross teach the transformer with a CNN network significantly improves the baseline method. Masked domain learning trains the transformer with extra supervision signal, which significantly boosts the performance. The experiment result empirically confirms our previous analysis that **even more supervision signal is needed to effectively train the transformer**. The addition of adversarial masked domain adaptation to reduce the domain gap between the original and masked domain further improves the model performance. The experiment result reveals that **the domain gap between the original and masked domain can negatively affect semi-supervised learning performance and reducing the domain gap helps to boost semi-supervised learning performance**. Finally, the extension of adversarial masked image modeling to CNN network provides extra improvement.

Comparison with SoTA Methods. In Table 2, we present the comparison results of our method with state-of-the-art methods on different label partitions across three public datasets. As can be observed, for all label partitions and datasets, our method outperforms existing methods significantly. For **ACDC**, our method outperforms the previous best method M-CnT by **10.1%** in Dice and **8.7mm** in HD on 3% partition, which is a tremendous improvement and

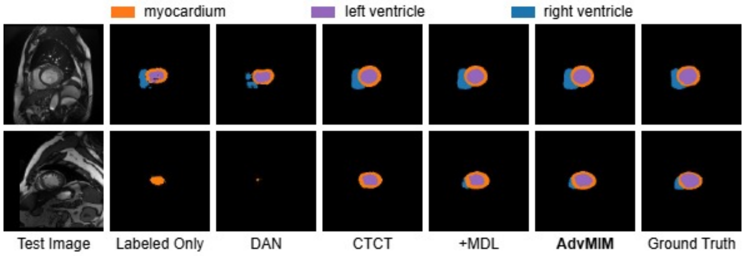

Fig. 2. Visual comparison with different methods on ACDC (3%).

Table 3. Effect of Mask Ratio on ACDC (3%) in Dice (%) and HD (mm).

mask ratio	0.1	0.3	0.5	0.7	0.9
Dice	84.7	85.0	85.1	**85.4**	80.5
HD	2.4	2.3	2.1	**2.0**	3.8

Table 4. Sensitivity analysis of λ_{adv} on ACDC (3%) in Dice (%) and HD (mm).

λ_{adv}	0.0001	0.001	0.01	0.1	1.0
Dice	85.0	**85.4**	85.3	85.3	76.6
HD	2.3	**2.0**	2.1	2.1	4.8

highlights the effectiveness of our method in handling limited labeled data scenario. For **Synapse**, our method significantly outperforms previous best method by **1.0%** and **3.4%** in Dice, **9.9mm** and **15.7mm** in HD on 15% and 30% partitions respectively. For **ISIC**, our method outperforms previous best method by **1.9%** in Dice and **9.9mm** in HD on 3% partition, where a tremendous improvement is observed in HD score. We further observe that our method maintains strong performance even with a small label partition, where other methods fail. Specifically, our method achieves **85.4%** and **79.8%** in Dice for ACDC and ISIC with only 3% labeled data, which highlights its remarkable applicability in real-world scenarios with extremely limited annotations.

Visualization Results. In Fig. 2, we present the visual comparison results of our method with different comparison methods. As shown in the figure, our method produces qualitatively much better segmentation masks when compared to existing adversarial training based method DAN [28], state-of-the-art method CTCT [15], and our ablated masked domain learning method.

Effectiveness of Mask Ratio. In Table 3, we present the effect of mask ratio on our method. Mask ratio controls the domain gap between masked and original domain. Too small mask ratio results in mask domain too similar to the original domain, which limits the extra supervision signal. Too large mask ratio increases the domain gap, which degrades the performance as supported by our Theorem 1. The default mask ratio of 0.7 gives the best performance.

Sensitivity Analysis. In Table 4, we present the sensitivity analysis of our method on the hyperparameter λ_{adv}. Experiment results show that our method is robust to the change of λ_{adv} in a wide range but too large value leads to poorer performance. The default value of $\lambda_{adv} = 0.001$ gives the best performance.

4 Conclusion

In this paper, we propose an adversarial masked image modeling method to fully unleash the potential of transformer for semi-supervised medical image segmentation. Our key contributions include the construction of a masked domain with masked image modeling for effective training of transformer with extra supervision signal and a theoretical analysis showing that the domain gap between masked and original domain can negatively affect semi-supervised learning performance. Thus, we propose a novel adversarial masked domain adaptation loss to minimize the domain gap. We also extend adversarial masked image modeling to CNN network. Extensive experiments on three public medical image segmentation datasets demonstrate the effectiveness of our method, where our method outperforms existing methods significantly.

Acknowledgments. This work was supported by the Agency for Science, Technology, and Research (A*STAR) through its IEO Decentralised GAP Under Project I24D1AG085.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. In: MICCAI 2017. Springer (2017)
2. Basak, H., Bhattacharya, R., Hussain, R., Chatterjee, A.: An exceedingly simple consistency regularization method for semi-supervised medical image segmentation. In: ISBI 2022. pp. 1–4. IEEE (2022)
3. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**, 151–175 (2010)
4. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., 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)
5. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR. pp. 2613–2622 (2021)
8. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: 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). In: ISBI 2018. pp. 168–172. IEEE (2018)

9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
11. Hu, X., Guo, R., Chen, J., Li, H., Waldmannstetter, D., Zhao, Y., Li, B., Shi, K., Menze, B.: Coarse-to-fine adversarial networks and zone-based uncertainty analysis for nk/t-cell lymphoma segmentation in ct/pet images. IEEE journal of biomedical and health informatics **24**(9), 2599–2608 (2020)
12. Huang, H., Huang, Y., Xie, S., Lin, L., Tong, R., Chen, Y.W., Li, Y., Zheng, Y.: Combinatorial cnn-transformer learning with manifold constraints for semi-supervised medical image segmentation. In: AAAI 2024. vol. 38, pp. 2330–2338 (2024)
13. Landman, B., Xu, Z., Igelsias, J.E., Styner, M., Langerak, T., Klein, A.: Segmentation outside the cranial vault challenge. In: MICCAI: multi Atlas labeling beyond cranial vault-workshop challenge (2015)
14. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896. Atlanta (2013)
15. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: International conference on medical imaging with deep learning. pp. 820–833. PMLR (2022)
16. Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: MICCAI 2021. pp. 318–329. Springer (2021)
17. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2017)
18. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: CVPR. pp. 12674–12684 (2020)
19. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: Proceedings of the european conference on computer vision (eccv). pp. 135–152 (2018)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI 2015. pp. 234–241. Springer (2015)
21. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
22. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. arxiv 2020. arXiv preprint arXiv:2012.12877 **2**(3) (2020)
23. Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. Neural Networks **145**, 90–106 (2022)
24. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2517–2526 (2019)

25. Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: MICCAI 2022. pp. 34–43. Springer (2022)
26. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
27. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: MICCAI 2019. pp. 605–613. Springer (2019)
28. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: MICCAI 2017. pp. 408–416. Springer (2017)
29. Zhu, L., Yang, K., Zhang, M., Chan, L.L., Ng, T.K., Ooi, B.C.: Semi-supervised unpaired multi-modal learning for label-efficient medical image segmentation. In: MICCAI 2021. pp. 394–404. Springer (2021)