SelfReg-UNet: Self-Regularized UNet for Medical Image Segmentation

Wenhui Zhu[1*†], Xiwen Chen[2*], Peijie Qiu[3*], Mohammad Farazi[1], Aristeidis Sotiras[3], Abolfazl Razi[2], and Yalin Wang[1]

[1] School of Computing and Augmented Intelligence, Arizona State University, AZ, USA
[2] School of Computing, Clemson University, SC, USA
[3] Washington University School of Medicine in St. Louis, MO, USA

Abstract. Since its introduction, UNet has been leading a variety of medical image segmentation tasks. Although numerous follow-up studies have also been dedicated to improving the performance of standard UNet, few have conducted in-depth analyses of the underlying interest pattern of UNet in medical image segmentation. In this paper, we explore the patterns learned in a UNet and observe two important factors that potentially affect its performance: (i) irrelative feature learned caused by asymmetric supervision; (ii) feature redundancy in the feature map. To this end, we propose to balance the supervision between encoder and decoder and reduce the redundant information in the UNet. Specifically, we use the feature map that contains the most semantic information (i.e., the last layer of the decoder) to provide additional supervision to other blocks to provide additional supervision and reduce feature redundancy by leveraging feature distillation. The proposed method can be easily integrated into existing UNet architecture in a plug-and-play fashion with negligible computational cost. The experimental results suggest that the proposed method consistently improves the performance of standard UNets on four medical image segmentation datasets. The code is available at https://github.com/ChongQingNoSubway/SelfReg-UNet

Keywords: Image Segmentation · UNet · Interpretability analysis

1 Introduction

Medical image segmentation is a pivotal application for computer-aided diagnosis and image-guided systems. Recently, deep learning has ascended as the leading method in this domain, primarily attributed to the landmark contribution of UNet [19]. UNet defines a generic segmentation network architecture by leveraging the encoder to project semantic information into low-level features and the decoder to progressively upsample semantic features to segmentation masks. Many of its follow-up works [4,23] have expanded this idea within the context of

---

* These authors contributed equally to this paper.
† Corresponding author: wzhu59@asu.edu

convolutional neural networks (CNN). Recently, numerous studies [3,1,27,22,17] have introduced vision transformer (ViT) [7] to address the limitations of CNN by using a self-attention mechanism. Although ViT has a large receptive field and captures long-range dependencies between different image patches, it struggles to preserve fine-grained local context due to the lack of locality. To mitigate this problem, methods [8,18] that bridge the gap between CNNs and ViTs (i.e., hybrid models) have been introduced in UNet design. Note that these methods also bring much computational complexity and the number of parameters. Over-parameterization is a common issue in deep learning, often leading to feature redundancy and poor feature representation [6,12,13]. However, this issue has not been formally investigated or considered by previous medical segmentation models.

Beyond the aforementioned methods, some efforts focus on optimizing the architectural structures of the UNet. Following this vein, Att-UNet [15] proposes an attention-based skip connection to suppress irrelevant features. UNet++ [31] replaces the standard skip connections (i.e., concatenation/addition) with the nested dense skip pathways. UCTransNet [24] thoroughly analyzes the effect of different skip connections and proposes a channel transformer to replace the conventional skip connection. These methods only investigate the information flow from encoder to decoder by manipulating the skip-connection. However, none of them have explored how to inform the encoder effectively via the learned features in the decoder, while our investigation reveals this information flow deserves more attention in a UNet. This is mainly because the decoder receives more supervision than the encoder, which provides a natural way to filter out irrelevant information.

In this paper, we perform empirical studies in two representative UNets (i.e., standard UNet [19] and SwinUNet [1]). Our analyses reveal two key findings: (i) Redundant features exist in the feature channel, with the shallow channels exhibiting more diversity than deep channels in a feature map; (ii) Asymmetric supervision between the encoder and the decoder in a UNet leads to semantic loss. This phenomenon diverges from observed trends in other computer vision tasks [29,26], where deep features exhibit to be more discriminative and can better localize the target of interest. To mitigate those issues, we introduce semantic consistency regularization and internal feature distillation to address semantic loss from asymmetric supervision and feature redundancy, respectively. This involves using more accurate semantics to supervise the other blocks and distilling information from shallow to deep channels in the feature map.

Contributions: **(i)** Our exploration uncover asymmetric supervision and feature redundancy in UNet, suggesting a novel direction for future model design. **(ii)** We suggest an orthogonal way to help the UNet discard irrelevant information and better preserve semantics by proposing a symmetric supervision regularization mechanism and leveraging the feature distillation. **(iii)** The proposed methods can be seamlessly integrated into existing UNet frameworks (e.g., CNN-UNet and ViT-UNet), offering performance gains with minimal extra cost.
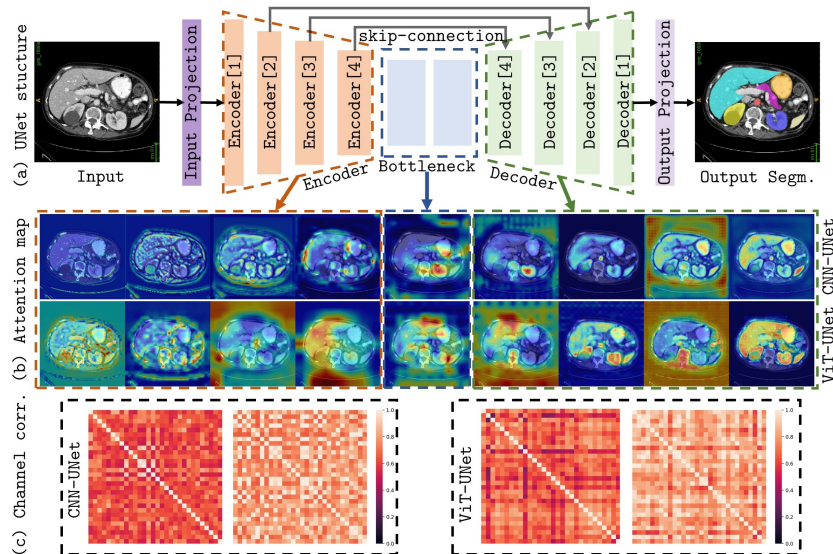
Fig. 1. (a)UNet structure. (b) The attention map in ViT/CNN-based UNet (For more examples, refer to Appendix A) (c) ViT/CNN-based UNet feature similarity matrix between shallow (Left) and deeper channel (Right).

2 Method

Preliminary. In this paper, we take a standard UNet defined in [19] with a depth of 5 as an example. Without loss of generality, we consider a unified UNet structure for both CNN-based UNet [19] and ViT-based UNet [1] in our investigation, which consists of an input and output projection block, 4 encoder and decoder blocks, as well as a bottleneck (see Fig. 1(a)). For this purpose, we first define the patch embedding block in a ViT-UNet as the combination of the input projection block and the first encoder block. We then define the last up-sampling block in a ViT-UNet (i.e., last patch expanding in SwinUNet [1]) as the last decoder block. Each encoder/decoder block of the defined UNet comprises $L = 2$ consective convolution/transformer blocks. The output projection block is a convolutional layer that maps the last feature map to a segmentation mask. For notation, we use $E_m^{(l)}$/ $D_m^{(l)}$ and $B^{(l)}$ to denote the l-th layer of the m-th encoder/decoder block and bottleneck, respectively. Accordingly, the corresponding feature map is denoted as F_m^l following order from encoder to decoder $(E_1^{(1)}, E_1^{(2)}, ..B^{(1)}.., D_2^{(2)}, D_1^{(1)})$.

2.1 Analysis on features learned in a UNet

We conduct analyses on features learned in UNets by employing two commonly used techniques: (i) gradient-weighted class activation mapping (Grad-CAM) [20]; (ii) similarity analysis in a feature map.

Asymmetric supervision in UNet. We observed two interesting phenomena as evident in Fig. 1(b): (i) The learning patterns exhibit an asymmetry between
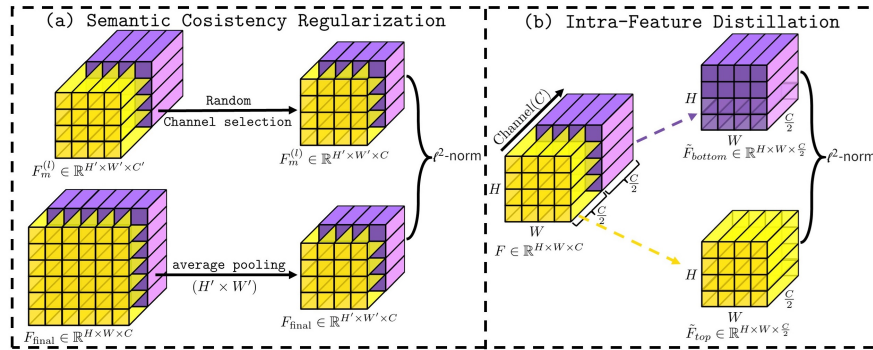
Fig. 2. Demostrating the operation based on feature for (a) semantic consistency regularization and (b) internal feature distillation.

the encoder and the decoder. The decoder could approximately locate some ground truth segmentation regions, and the encoder tends to capture unrelated information (E_3, E_4), dispersing the interest of patterns towards the boundaries. (ii) In the decoder, the block (D_1) located further to the end demonstrates an accurate understanding of ground truth segmentation. Meahwhile, the blocks (D_2,D_4) learned the irrelevant information. The main reason lies in the different intensity of supervision signals each block receives. When tracing back from the output projection directly supervised by the ground truth, the supervision signal progressively diminishes. It leads to semantic loss, with some blocks (e.g., E_1,E_3,E_4,B) in the encoder even activating regions unrelated to segmentation.

Redundant Features in UNet. Empirical investigations have shown that overparameterized CNN/ViT models tend to learn redundant features, leading to poor visual concepts [6,12,13]. Taking the output of E_1 as an illustration, we calculate the feature similarity matrices in the channel dimension at both the shallow and the deep levels. As depicted in Fig. 1(c), we observe two phenomena in both ViT/CNN-based UNets: (i) Feature redundancy is prevalent in deep layers, with a high similarity matrix indicating the learning of similar features across channels. (ii) The shallow layers exhibit significant diversity, evidenced by a low similarity matrix. The over-parameterization that exists in UNet, akin to that in other networks, underlies these phenomena. The resultant redundant features often come with task-irrelevant visual features, leading to performance degradation and unnecessary computation overhead.

2.2   Solutions

Semantic Consistency Regularization. Various studies have been proposed to tackle the loss of semantics in natural images by employing strategies such as knowledge distillation and feature alignment, with the goal of leveraging accurate features to guide those less informative ones [2,13,9,5,28]. Inspired by these works in natural images, we propose to use the feature map that contains the most semantic information (i.e., D_1 as observed) to provide additional supervision to

the rest blocks in a UNet. As a result, we propose a generic paradigm, termed semantic consistency regularization (SCR), to balance the supervision between the encoder and the decoder. For simplicity, we demonstrate our idea using the feature distillation mechanism in [2,13] due to its popularity and simplicity (given as a mean square loss); while other knowledge distillation methods can be alternatives (given as KL divergence). To this end, we define the proposed SCR as

$$\mathcal{L}_{SCR} = \frac{1}{|M-1||\mathcal{I}|} \sum_{m=1}^{M-1} \sum_{I \sim \mathcal{I}} \| \operatorname{RCS}(F_m^i) - \operatorname{AvgPool}(F_{final}) \|^2,$$

Where F_{final} is the feature map located at the last decoder block ($\mathtt{D}_1$), and F_i^m is all the feature map situated in the ith layer of the mth block ($\mathtt{E}_1^{(1)}$, $\mathtt{E}_1^{(2)}$,..., $\mathtt{D}_3^{(2)}$,$\mathtt{D}_2^{(1)}$,$\mathtt{D}_2^{(2)}$) except the $\mathtt{D}_1$. To align features in channel and spatial dimensional, we employed an average-pooling and random channel selection operation (RSC) as shown in Fig 2.(a). It is worth noting that the channel selection does not introduce extra modules [9,5,28], reducing computation and semantic conflicts. The L_2 norm is used as a distance metric.

Internal Feature Distillation. To solve the feature redundancy problem, some channel shrinkage methods have been proposed related to fields of model filter pruning, which leverage the L_p norm penalty to induce a sparsity prior on channel salience [30,25,14]. Inspired by this, we employed the L_p norm for information distillation from shallow (top-half channel feature) to deep (bottom-half channel one), which guides the deeper features learned the useful context information. It can be formulated as:

$$\mathcal{L}_{IFD} = \frac{1}{|M||\mathcal{I}|} \sum_{m=1}^{M} \sum_{I \sim \mathcal{I}} \| \widetilde{F_m^i} - \overline{F_m^i} \|^p,$$

where F_i^k denotes all feature maps situated in the ith layer of the mth block ($\mathtt{E}_1^{(1)}$, $\mathtt{E}_1^{(2)}$,..., $\mathtt{D}_2^{(1)}$,$\mathtt{D}_2^{(2)}$,$\mathtt{D}_1^{(1)}$), $\widetilde{F}$ is the deep channel feature, and $\overline{F}$ was shallow channel feature. As shown in Fig 2.(b), we partitioned the channels into top and bottom halves, using this division as boundaries to ensure the same number of features in both shallow and deep. Following [13,25,30], we employed the L_2 norm. Compared with methods that introduced extra modules for reducing redundancy [16,11], the $\mathcal{L}_{IFD}$ is simple and cost-free.

Objective Function. The total loss is weighted of $\mathcal{L}_{SCR}$ and $\mathcal{L}_{IFD}$ with the standard combination of cross-entropy and dice losses $\mathcal{L}_{cd}$ [24,8,1,19] that is evaluated between the prediction and the ground truth segmentation results.

$$\mathcal{L} = \mathcal{L}_{cd} + \lambda_1 \mathcal{L}_{SCR} + \lambda_2 \mathcal{L}_{IFD},$$

where λ_1 and λ_2 are balance parameters.

3   Experiments and Results

3.1   Dataset

Synapse Multi-Organ Segmentation. Following [1,3,8], the Synapse dataset comprises 30 cases comprising 3779 axial abdominal clinical CT images. Of these, 18 samples are designated for training, while 12 are reserved for testing. We evaluate performance on eight abdominal organs using the average Dice Similarity Coefficient (DSC) as the evaluation metric.
Automated cardiac diagnosis challenge dataset. The ACDC dataset comprises 100 cardiac MRI scans from diverse patients, each labeled with left ventricle (LV), right ventricle (RV), and myocardium (MYO). Following [3,18], we allocate 70 cases (1930 axial slices) for training, 10 for validation, and 20 for testing. We evaluate our method using the DSC as the evaluation metric.
Nuclear segmentation and Gland segmentation. The Gland segmentation dataset (GlaS) [21] has 85 images for training and 80 for testing. The Multi-Organ Nucleus Segmentation (MoNuSeg) dataset [10] has 30 images for training and 14 for testing. Following [24], we perform the three times 5-fold cross-validation results on GlaS and MoNuSeg datasets. The average DSC and IoU are used as evaluation metrics.
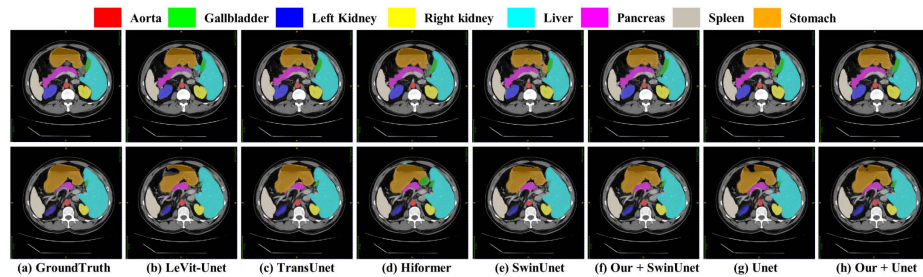

Fig. 3. Comparison of segmentation performance in Synapse dataset.

3.2   Experiment settings

We evaluate the effectiveness of our proposed loss on both SwinUNet and UNet in these four datasets, and the training setups (i.e., batch size, optimizer, learning rate, etc.) are consistent with [1]. All experiments were conducted with an input image size of 224 x 224 and the same data augmentation and preprocessing in [8,18,24], using an Nvidia GTX3090 with 24GB of memory for training. Following in [1,19], pre-trained weights from ImageNet were employed in SwinUNet, while UNet was trained from scratch.
Comparison with state-of-the-art methods. We compare our methods with recent SOTA models, including R50 UNet [3], Att-UNet [15], UNet++ [31], TransUNet [3],swinUNet [1], LeViT-UNet [27], DeepLabv3 [4], HiFormer [8], PVT-cascade [18], UCTransNet [24], MedT [22].

Table 1. Comparison with SOTA methods on Synapse multi-organ CT dataset. Δ denotes the improvement gain (%) by comparing with the original method.

Methods	Average DSC	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
R50 U-Net	74.68	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
U-Net	76.85	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
R50 Att-UNet	75.57	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
Att-UNet	77.77	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
TransUNet	77.48	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUNet	79.13	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
LeViT-UNet	78.53	78.53	62.23	84.61	80.25	93.11	59.07	88.86	72.76
DeepLabv3	77.63	88.04	66.51	82.76	74.21	91.23	58.32	87.43	73.53
HiFormer	80.29	85.63	73.29	82.39	64.84	94.22	60.84	91.03	78.07
Our + UNet	80.34	88.74	71.78	85.32	80.71	93.80	62.22	84.78	75.39
Δ	+3.49	-0.33	+2.06	+7.55	+12.11	+0.37	+8.24	−1.89	−0.19
Our + SwinUNet	**80.54**	86.07	69.65	85.12	82.58	94.18	61.08	87.42	78.22
Δ	+1.41	+0.60	+3.12	+1.84	+2.97	−0.11	+4.50	−3.24	+1.62

Table 2. Comparison of different methods in ACDC dataset.

Methods	Avg DSC	RV	Myo	LV
R50 + AttnUNet	86.75	87.58	79.2	93.47
ViT + CUP	81.45	81.46	70.71	92.18
UNet	89.68	87.17	87.21	94.68
TransUNet	89.71	86.67	87.27	95.18
SwinUNet	88.07	85.77	84.42	94.03
LeViT-UNet	88.21	85.56	84.75	94.32
Hiformer	90.82	88.55	88.44	95.47
PVT - CASCADE	90.45	87.20	88.96	95.19
Our + UNet	91.43	88.92	89.49	95.88
Δ	+1.75	+1.75	+2.28	+1.20
Our + SwinUNet	**91.49**	89.49	89.27	95.70
Δ	+3.42	+3.72	+4.85	+1.67

Table 3. Comparison of different methods in Glas and MoNuSeg datasets.

Method	Glas		MoNuSeg	
	DSC (%)	IOU (%)	DSC (%)	IOU (%)
U-Net	85.45±1.25	74.78±1.67	76.45±2.62	62.86±3.00
UNet++	87.56±1.17	79.13±1.70	77.01±2.10	63.04±2.54
AttUNet	88.80±1.07	80.69±1.66	76.67±1.06	63.47±1.16
MRUNet	88.73±1.17	80.89±1.67	78.22±2.47	64.83±2.87
TransUNet	88.40±0.74	80.40±1.04	78.53±1.06	65.05±1.28
MedT	85.92±2.93	75.47±3.46	77.46±2.38	63.37±3.11
SwimUNet	89.58±0.57	82.06±0.73	77.69±0.94	63.77±1.15
UCTransNet	90.18±0.71	82.96±1.06	79.08±0.67	65.50±0.91
Our + UNet	90.93±0.20	84.08±0.41	**80.18±0.19**	**67.07±0.29**
Δ	+5.48	+9.3	+3.73	+4.21
Our + SwinUNet	**91.62±0.16**	**85.29±0.30**	79.38±0.15	65.87±0.21
Δ	+2.04	+3.23	+1.69	+2.10

3.3 Results

The results are shown in Table 1 for Synapse dataset, Table 2 for ACDC dataset, and Table 3 for Glas and MoNuSeg datasets. The main observation is our proposed loss is effective and can lead to substantial gains. Specifically, by utilizing our proposed loss, UNet obtains a 3.49%, 1.75%, 5.48%, and 3.73% improvement in average DSC in the four datasets, respectively. Likewise, in these four datasets, SwinUNet can obtain a 1.41%, 3.42%, 2.04%, and 1.69% DSC improvement.

(A). Raw Image   **(B). Ground Truth**   **(G). UCTransNet**   **(D). TransUNet**   **(E). MedT**   **(F). Unet**   **(G). Our + Unet**   **(H). SwinUnet**   **(I). Our + SwinUnet**
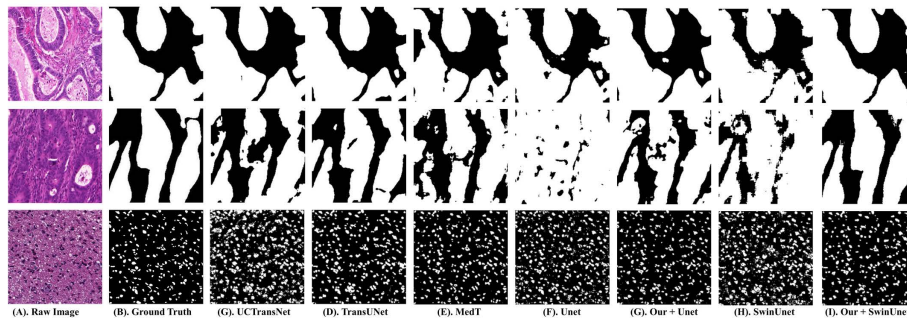
Fig. 4. Comparison of segmentation performance in Glas and MoNuSeg dataset.
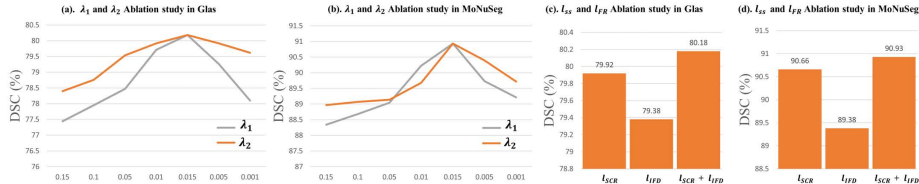
Fig. 5. Ablation study for balance parameters λ and loss based on Glas and MoNuSeg.

Consequently, our enhanced models can outperform the previous State-of-the-art (SOTA) methods. For example, our enhanced UNet and SwinUNet exhibit improvements of 0.05% and 0.25% in average DSC, respectively. Compared with Hiformer, TransUNet, UNet, and swinUNet [19,1,8,3], the performance gains of our enhanced models primarily derive from tackling challenging organ segmentation tasks such as the gallbladder, left kidney, right kidney and pancreas, as illustrated in Fig 3 and Table 1. Similarly, in ACDC dataset, compared with SOTA HiFormer [8], our UNet and swinUNet methods demonstrate enhancements of 0.6% and 0.67% in average DSC, respectively. This superiority is also generalized to Glas and MoNuSeg datasets, and our methods exhibit boosted performance over the previous SOTA method, UCTransNet, with 0.7% and 1.1% DSC gain in both datasets, respectively. Fig 4 demonstrates our method's notable improvement in Glas dataset. Note that UNet with optimized skip connections (e.g., UCTTransNet [24]) leads to irrelevant segmentation and incomplete shapes. Especially segments resembling the background (last row). While SwinUNet, with our proposed loss, demonstrates results that are closely comparable to the ground truth, featuring full shapes and clear backgrounds, particularly in the hard to improvement sample (the third row in Fig 4). These observations support that our approach can provide contextual supervision, ensuring correct semantics across blocks and preventing the learning of irrelevant features.

3.4   Ablation studies

To look deeper into our method, we perform a series of ablation studies using UNet on the Glas and MoNuSeg dataset. The results and analysis are as follows:

Loss balance hyperparameters. As result shown in Fig 5(a) and (b), we observe that the $\lambda_1 = 0.015$ and $\lambda_2 = 0.015$ is the optimal setting. As the weight of λ_1 decreases (l_{SCR}), the performance deteriorates rapidly. This indicates that the lack of correct semantic supervision is a reason for the performance decline.

Effectiveness of the proposed losses. We conducted the ablation study with the different losses as shown in Fig 5(c) and (d). Both datasets exhibit consistent results, achieving optimal performance when both losses are used simultaneously.

4   Conclusion

In this paper, we unveil the problem of asymmetric supervision and feature redundancy in UNet-based medical image segmentation, suggesting a novel approach through loss optimization that incorporates semantic consistency regularization and internal feature distillation. Our experimental results demonstrate that addressing these two issues indeed has the potential to improve ViT/CNN-based UNet models, and the proposed method holds potential across a wide range of medical image segmentation tasks. In the future, our findings can aid in the design of UNet, and we will explore additional solutions from this perspective.

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

References

1. 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)
2. Chen, D., Mei, J.P., Zhang, Y., Wang, C., Wang, Z., Feng, Y., Chen, C.: Cross-layer distillation with semantic calibration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 7028–7036 (2021)
3. 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)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
5. Chung, I., Park, S., Kim, J., Kwak, N.: Feature-map-level online adversarial knowledge distillation. In: International Conference on Machine Learning. pp. 2006–2015. PMLR (2020)
6. Dalvi, F., Sajjad, H., Durrani, N., Belinkov, Y.: Analyzing redundancy in pre-trained transformer models. arXiv preprint arXiv:2004.04010 (2020)
7. 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)
8. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6202–6212 (2023)
9. Hou, Y., Ma, Z., Liu, C., Loy, C.C.: Learning lightweight lane detection cnns by self attention distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1013–1021 (2019)

10. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology. IEEE Trans Med Imaging **36**(7), 1550–1560 (Jul 2017)
11. Li, G., Jin, D., Yu, Q., Qi, M.: Ib-transunet: Combining information bottleneck and transformer for medical image segmentation. Journal of King Saud University - Computer and Information Sciences **35**(3), 249–258 (2023)
12. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient convnets. In: International Conference on Learning Representations (2017)
13. Li, L.: Self-regulated feature learning via teacher-free feature distillation. In: European Conference on Computer Vision. pp. 347–363. Springer (2022)
14. Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., Zhang, C.: Learning efficient convolutional networks through network slimming. In: Proceedings of the IEEE international conference on computer vision. pp. 2736–2744 (2017)
15. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
16. Pan, S., Liu, X., Xie, N., Chong, Y.: Eg-transunet: a transformer-based u-net with enhanced and guided models for biomedical image segmentation. BMC bioinformatics **24**(1), 85 (2023)
17. Qiu, P., Yang, J., Kumar, S., Ghosh, S.S., Sotiras, A.: Agileformer: Spatially agile transformer unet for medical image segmentation. arXiv preprint arXiv:2404.00122 (2024)
18. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6222–6231 (2023)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
21. Sirinukunwattana, K., Pluim, J.P.W., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., hm, A., Ronneberger, O., Cheikh, B.B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D.R.J., Rajpoot, N.M.: Gland segmentation in colon histology images: The glas challenge contest. Med Image Anal **35**, 489–502 (Jan 2017)
22. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 36–46. Springer (2021)
23. WANG, F.k., HUANG, Y.q., HUANG, Z.c., Hao, S., HUANG, C., Xi, Q., QIAN, W.q.: Mrunet: A two-stage segmentation model for small insect targets in complex environments. Journal of Integrative Agriculture **22**(4), 1117–1130 (2023)
24. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2441–2449 (2022)
25. Wen, W., Wu, C., Wang, Y., Chen, Y., Li, H.: Learning structured sparsity in deep neural networks. Advances in neural information processing systems **29** (2016)

26. Wolchover, N.: New theory cracks open the black box of deep learning (2018)
27. Xu, G., Zhang, X., He, X., Wu, X.: Levit-unet: Make faster encoders with transformer for medical image segmentation. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 42–53. Springer (2023)
28. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
29. Zhang, D., Zhang, H., Tang, J., Hua, X.S., Sun, Q.: Self-regulation for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6953–6963 (2021)
30. Zhao, C., Ni, B., Zhang, J., Zhao, Q., Zhang, W., Tian, Q.: Variational convolutional neural network pruning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2780–2789 (2019)
31. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)