


Only Classification Head is Sufficient for Medical Image Segmentation

Hongbin Wei¹, Zhiwei Hu², Bo Chen², Zhilong Ji², Hongpeng Jia¹,
Lihe Zhang¹, , Huchuan Lu¹

¹ Dalian University of Technology, China

{weihongbin,22109042}@mail.dlut.edu.cn, {zhanglihe,lhchuan}@dlut.edu.cn,

² Tomorrow Advancing Life

{chenbo2,huzhiwei3,jizhilong}@tal.com,

Abstract. Medical image segmentation is a pivotal research domain that has garnered widespread attention in contemporary medical diagnostics. In pursuit of enhancing network efficacy, researchers have taken great efforts to develop various well-designed decoders. Unfortunately, due to the limited medical training data, the issues of underfitting and overfitting frequently arise. To this end, we undertake plentiful experiments to decouple the encoder and decoder components, and obtain a critical finding that excessively complex decoders impede the encoder’s potentiality of feature extraction. Inspired by some remarkable image generation work, we devise a straightforward segmentation network, which incorporates a pre-trained encoder backbone network and a pixel classification head. Our network not only ensures adequate feature decoding ability but also maximizes feature representation capability of the backbone. Experimental results on four datasets of three tasks show the outstanding performance against the state-of-the-art methods. The source code will be publicly available at <https://github.com/weihongbin/CHNet>

Keywords: Medical image segmentation · Encoder-decoder decoupling
· Classification head.

1 Introduction

Medical image segmentation aims to separate objects of interest from the surrounding environment in medical images, such as human tissues, organs, and pathological regions. It has broad applications in medical research, clinical diagnosis, pathological analysis, and assisted surgery, and is of significant importance.

Due to patient privacy concerns, acquiring medical images is difficult, and medical images need to be annotated by professional doctors, which results in expensive labelling cost. As a result, the scale of current medical image datasets is generally small. The lack of enough training data makes it difficult for neural networks to be sufficiently trained. Many natural image segmentation methods cannot be directly transferred to medical segmentation field, which severely limits its development.

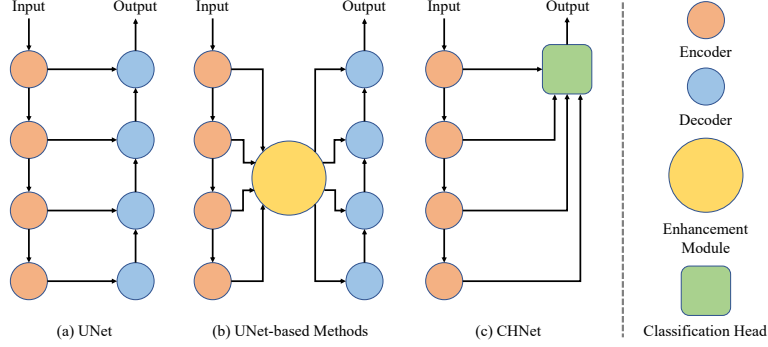


Fig. 1. Illustration of different medical image segmentation architectures

Currently, UNet [24] and its variants [19, 15, 8, 18, 17, 13] dominate medical segmentation field. They utilize the encoder-decoder architecture. As shown in Fig.1 (a), the UNet adopts a symmetrical encoder-decoder structure, which extracts and reconstructs features through multiple downsampling and upsampling operations. The encoder can obtain features of different scales, which go through upsampling and skip-connection operations into the decoder. These features are fused to obtain segmentation results. Researchers believe that this direct connection method is too rough to fully tap into the potential of features at different levels. Therefore, various feature enhancement modules have been designed to strengthen feature expression ability, as shown in Fig.1 (b). A large amount of research shows that using an encoder that has been pre-trained on a large-scale dataset such as ImageNet can significantly improve the performance of downstream tasks. Therefore, the common paradigm is to design innovative decoder based on a pre-trained encoder and then fine-tune the whole model on medical images for domain adaptation.

In an encoder-decoder structured neural network, the overall performance is determined by both the feature extraction ability of the encoder and the feature inferring ability of the decoder. In our study, we decouple the network to investigate the two components separately, and find that the encoders of identical structure, which are training with different decoders, exhibit different feature extraction capabilities, which can be attributed to two factors: 1) different inferring abilities of the decoders, and 2) the influence of the decoder on the feature extraction ability of the encoder during coupled training.

In image generation field, some studies [28, 29, 3] simultaneously generate images and their ground truths. The generation of ground truths is accomplished by using a dedicated pixel classification head. Specifically, the input of the classification head is the intermediate features of the generated image, and the output is supervised by a small number of ground truths. These approaches have achieved remarkable results, which indicates that the feature inferring ability of the pixel classification head is sufficient in scenarios where training data is limited.

Recently, the development of encoders has progressed from convolutional neural networks (CNNs) to transformers, which have increasingly powerful feature

extraction capabilities. Moreover, a plethora of pre-training strategies have been introduced to learn prior knowledge so that the extracted features already contain rich semantic information. We designed a simple classification head decoder, as shown in Fig.1 (c), which not only possesses sufficient feature inferring capabilities but also maximizes the feature extraction capabilities of the encoder. Our model has a simple and flexible structure, with few parameters and computational requirements. Most importantly, it achieves remarkably high performance. Overall, our contributions can be summarized as follows:

- We proposed a simple yet effective segmentation network (CHNet) with a classification head as the decoder, which can sufficiently learn representation capacity of the whole network on the limited training data.
- We decoupled the encoder-decoder framework, and experimentally analyzed the dependencies of feature representation between encoder and decoder. It was observed that the complex decoder design can suppress the encoding capability of the encoder.
- Extensive experiments show that the proposed model achieves state-of-the-art performance on four datasets of three tasks across multiple metrics.

2 Related Work

As an important method of medical image analysis, medical image segmentation aims to provide a more clear picture of changes in anatomical or pathological structures in an image, thus providing a reliable basis for clinical diagnosis and early diagnosis of diseases. It plays a crucial role in computer-aided diagnosis and intelligent medical treatment, greatly improving the efficiency and accuracy of diagnosis.

Feature extraction for medical images is more difficult than for normal RGB images, as the former often suffers from blur, noise and low contrast. Traditional medical image segmentation algorithms rely heavily on human prior knowledge and have insufficient generalization capabilities, making it difficult to obtain satisfactory results. With the rapid development of deep learning techniques in recent years, convolutional neural networks [32, 33, 21, 34, 30, 20, 31, 22] have successfully achieved hierarchical feature representation of images, which has become a hot research topic in computer vision.

U-Net [24] is widely used in medical image segmentation and has become the benchmark for most medical image segmentation tasks. U-Net uses skip connections to combine the high-level semantic feature maps from the decoder and corresponding low-level detailed feature maps from the encoder. It is widely believed that the success of U-Net depends on the U-shape structure, and many U-Net-based models have been proposed. Atten-UNet [19] embeds an attention gate in each transition layer between encoder and decoder blocks, which automatically learns to focus on target structures of different shapes and sizes. UNet++ [35] introduces nesting and dense skip connections to reduce the semantic gap between encoders and decoders. Although reasonable performance

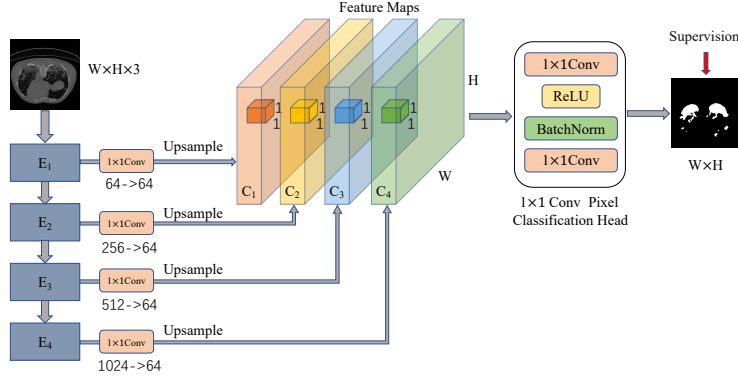


Fig. 2. The overall architecture of the proposed CHNet.

can be achieved, nested network structures are too complex to check for sufficient information at full scale. In each feature fusion, UNet3+ [13] aggregates feature maps at all scales using comprehensive skip connections to make more complete use of the full-scale feature information.

CNN can only capture local information, while transformer excels at direct global relationship modeling. Recently, the Transformer architecture has been successful in many tasks. Some works [12], [7] explore its effectiveness for the medical vision tasks. NTNet [12] is a simple but powerful hybrid transformer architecture, which integrates self-attention into CNN to enhance medical image segmentation. Another representative transformer-based model is TransUNet [7], which has the advantages of both transformers and U-Net. It encodes tokenized image patches from CNN feature maps into an input sequence to extract global context, and utilizes the low-level CNN features via a u-shaped hybrid architectural design.

3 The Proposed Method

3.1 Overall Architecture

Our design principles are simplicity and effectiveness, as illustrated in Fig.2. It is primarily composed of four encoder blocks and a single pixel classification head. Following the works [10, 34], Res2Net-50 was selected as the backbone for feature extraction, yielding four layers of feature representations. Additionally, to suit medical segmentation task, we make several modifications to the backbone. Specifically, we remove the last encoder block of Res2Net-50, as it does not provide any discernible performance gain. Furthermore, to minimize computational complexity, the 1×1 convolutional layers are employed to reduce the feature channels. Subsequently, these features are directly upsampled to the original image resolution and concatenated along the channel dimension to form a $256 \times W \times H$ feature map. This design effectively mitigates semantic information loss and boosts the representation power of modal.

In addition, a classification head was implemented for pixel-level classification, which consists of 1×1 convolutional layers, ReLU non-linear activation, and BatchNorm layers. The classification head directly conducts pixel classification along channel direction, thereby generating a complete prediction map.

3.2 Loss Function

The total loss function can be formulated as follows:

$$L_{total} = L_{IoU}^w + L_{BCE}^w, \quad (1)$$

where L_{IoU}^w and L_{BCE}^w represent the weighted IoU [27] loss and binary cross entropy (BCE) [27] loss, which restrict the prediction map in terms of the global structure and local details. They have been widely adopted in segmentation tasks. We use the same definitions as in [10, 23] and their effectiveness has been validated in these works.

4 Encoder-Decoder Decoupling Analyses

Typically, the encoder and decoder of a model are trained jointly in an end-to-end manner, without attention to their individual performance or the coupling effect. To investigate this issue, we conduct a series of decoupling experiments using U-Net [24], Atten-UNet [19], and the proposed CHNet. To ensure fairness, the backbone architectures of the three networks are identical, and the only variation is the decoder. Firstly, the three networks are trained to reach the optimal performance, respectively. And then the models are decoupled to obtain three backbones of different parameters. After freezing each backbone, different decoders are separately connected for decoupling research. Here, all reported results are conducted on the COVID-19 Lung dataset [2, 1]. They also have the same tendencies on the other datasets and tasks.

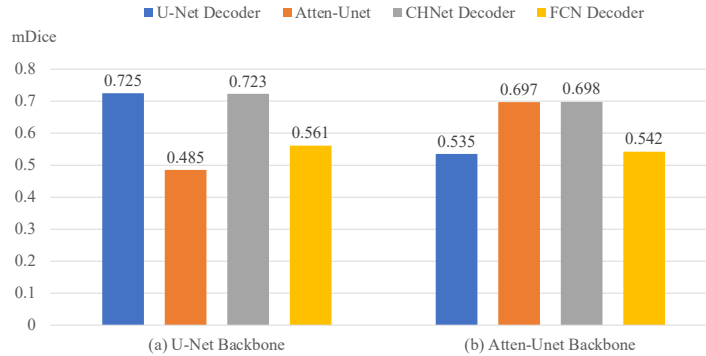


Fig. 3. Feature Inferring capability of different decoders.

4.1 Feature Inferring Capability of Decoder

The backbones of U-Net and Atten-UNet are combined with the decoders of CHNet and FCN-32s [16] to form new networks, respectively. After the backbones are all frozen, we retrain the decoders, which are randomly initialized as done in their original end-to-end learnt models. As shown in Fig.3 (a), the CHNet decoder and U-Net decoder perform similarly based on the U-Net backbone. Meanwhile, our decoder and Atten-UNet decoder also behave the same way in Fig.3 (b). These results indicate that the decoder of classification head has similar feature inferring capability to the complex U-Net and Atten-UNet decoders. The poor performance of FCN decoder, Atten-UNet decoder in Fig.3 (a) and U-Net decoder in Fig.3 (b) indicates the bad fitting ability of these decoders. In addition, when these decoders load the parameters of the optimal end-to-end models as initialization, and then are further trained based on the frozen backbone, the results are just the same.

4.2 Feature Extraction Capability of Encoder

In order to evaluate the encoder, we choose a simplest decoder FCN-32s [16] and combine it with the backbones of CHNet, U-Net and Atten-UNet to form new networks, respectively. Similarly, all the backbones are frozen and we retrain the decoder. Because this decoder is very light, its own influence on the final performance can be ignored. Thus, the prediction results mainly reflect the abilities of different backbones. From Fig.4 (a), we can see that the extraction ability of CHNet backbone is much better than those of U-Net and Atten-UNet, which actually reflects that the complex designs of U-Net and Atten-UNet decoders suppress their own backbones. In addition, we implement the same experiments on the Atten-UNet decoder, as shown in Fig.4 (b). Here, the original backbone only loads ImageNet pre-trained parameters. The relatively complex decoder is able to well fit the gaps between the backbones (even though the backbone does not see medical image data) and achieves similar results except the self-coupled backbone. The results in Fig.4 (b) also mean that a complexly structured decoder is more easy to deeply couple its own backbone.

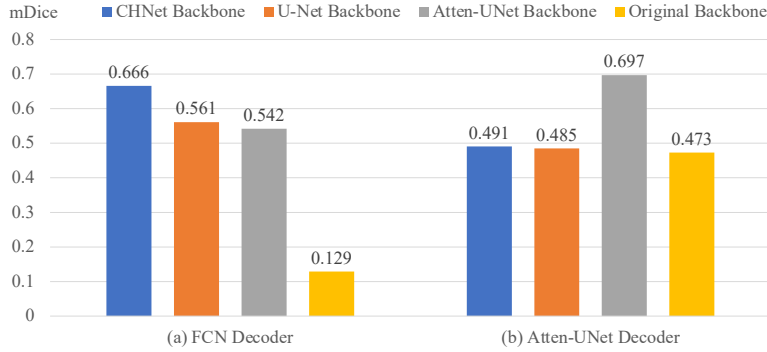


Fig. 4. Feature extraction capability of different encoders.

5 Experiments

5.1 Datasets

We verify the effectiveness of the proposed framework on three medical segmentation tasks, which cover diverse data modalities, such as computed tomography (CT), ultrasound imaging, and color colonoscopy imaging.

COVID-19 Lung Infection. Few publicly available COVID-19 lung CT datasets are suitable for infection segmentation. To have relatively sufficient samples for training, we merged two datasets [2, 1] to obtain 1,277 high-quality CT images. We divide them into 894 training images and 383 testing ones.

Breast Ultrasound Segmentation. The BUSI [25] is a common dataset for breast ultrasound segmentation, which consists of 780 images acquired from 600 female patients, including 133 normal cases, 437 benign tumors, and 210 malignant tumors. We compared with two well-established task-specific methods [5, 6] and performed a four-fold cross-validation on this dataset.

Polyp Segmentation. Two benchmark datasets CliniCDB [4] and Kvasir [14] are used. We adopt the same training set as [10], that is, 550 samples from the ClinicDB and 900 samples from the Kvasir are used for training. And the remaining 62 images and 100 images are used for testing.

5.2 Evaluation Metrics

There are many popular metrics used in different medical segmentation branches. Following [11], five metrics are employed for quantitative evaluation, including mean Dice ($mDice$), *Precision*, *Recall*, S-measure (S_α) and mean absolute error (MAE). Following [6], *Jaccard*, *Precision*, *Recall*, and *Dice* are more commonly used for breast tumor segmentation. For polyp segmentation, mean Dice ($mDice$), mean IoU ($mIoU$), the weighted F-measure (F_β^ω), S-measure (S_α), E-measure (E_ϕ^{max}) and mean absolute error (MAE) are widely used.

5.3 Implementation Details

Our model is implemented based on the PyTorch framework and trained on a single 3090 GPU with mini-batch size 16. We resize the inputs to 352×352 . Random horizontally flipping and random rotate augmentation are used.

For the optimizer, we adopt the SGD, and the momentum and weight decay are set as 0.9 and 0.0005, respectively. Warm-up and linear decay strategies are used to adjust the learning rate. For any medical image sub-tasks, the above training strategy is same. The difference among these models is only in the number of training epochs due to different convergence speeds. Specifically, the number of training epochs settings in the polyp segmentation, breast tumor segmentation and COVID-19 Lung Infection are 50, 100 and 200, respectively.

Table 1. Quantitative comparisons on the COVID-19 Lung dataset. Top 2 scores are highlighted in red and blue, respectively. “†” represents the medicine-specific method.

Methods	Backbone	$mDice \uparrow$	$Precision \uparrow$	$Recall \uparrow$	$S_\alpha \uparrow$	$MAE \downarrow$
U-Net[24]	R2-50	0.725	0.744	0.810	0.819	0.010
Atten-UNet[19]	R2-50	0.697	0.720	0.803	0.799	0.013
UTNet[12]	R-50 + ViT-B16	0.735	0.782	0.786	0.836	0.007
TransUnet[7]	R-50 + ViT-B16	0.710	0.770	0.776	0.831	0.007
Inf-Net [†] [11]	R2-50	0.783	0.774	0.852	0.843	0.007
BCS-Net [†] [9]	R2-50	0.763	0.775	0.763	0.840	0.007
Ours	R2-50	0.800	0.816	0.830	0.846	0.006

Table 2. Quantitative comparisons on the breast ultrasound dataset.

Methods	Backbone	$Jaccard \uparrow$	$Precision \uparrow$	$Recall \uparrow$	$Dice \uparrow$
U-Net[24]	R2-50	65.73±1.49	81.25±1.00	74.53±1.95	74.65±1.29
Atten-UNet[19]	R2-50	65.70±0.88	79.27±1.60	77.03±1.05	75.06±1.14
UTNet[12]	R-50 + ViT-B16	67.46±1.78	79.88±1.22	74.82±1.95	74.41±1.39
TransUnet[7]	R-50 + ViT-B16	71.47±0.98	81.66±1.52	80.78±1.63	79.00±0.79
SKU-Net [†] [5]	SKs	64.48±2.37	75.37±3.22	78.56±3.27	74.03±2.21
NU-Net [†] [6]	Deeper U-Net	68.86±1.99	78.90±2.26	82.48±2.14	77.79±1.88
Ours	R2-50	72.47±0.62	81.93±0.38	82.94±0.95	80.30±0.57

Table 3. Quantitative comparisons on the polyp segmentation dataset ClinicDB.

Methods	Backbone	$mDice \uparrow$	$mIoU \uparrow$	$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	$E_\phi^{max} \uparrow$	$MAE \downarrow$
U-Net[24]	R2-50	0.890	0.839	0.877	0.920	0.954	0.011
Atten-UNet[19]	R2-50	0.909	0.864	0.899	0.937	0.962	0.009
UTNet[12]	R-50 + ViT-B16	0.860	0.818	0.856	0.910	0.963	0.017
TransUnet[7]	R-50 + ViT-B16	0.847	0.798	0.831	0.907	0.920	0.020
PraNet [†] [10]	R2-50	0.899	0.849	0.896	0.936	0.963	0.009
SANet [†] [26]	R2-50	0.916	0.859	0.909	0.939	0.971	0.012
Ours	R2-50	0.926	0.882	0.915	0.946	0.969	0.009

Table 4. Quantitative comparisons on the polyp segmentation dataset Kvasir.

Methods	Backbone	$mDice \uparrow$	$mIoU \uparrow$	$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	$E_\phi^{max} \uparrow$	$MAE \downarrow$
U-Net[24]	R2-50	0.855	0.790	0.830	0.880	0.918	0.042
Atten-UNet[19]	R2-50	0.883	0.823	0.861	0.900	0.931	0.037
UTNet[12]	R-50 + ViT-B16	0.862	0.803	0.843	0.886	0.911	0.042
TransUnet[7]	R-50 + ViT-B16	0.869	0.816	0.847	0.899	0.920	0.040
PraNet [†] [10]	R2-50	0.898	0.840	0.885	0.915	0.944	0.030
SANet [†] [26]	R2-50	0.904	0.847	0.892	0.915	0.949	0.028
Ours	R2-50	0.911	0.864	0.900	0.924	0.950	0.026

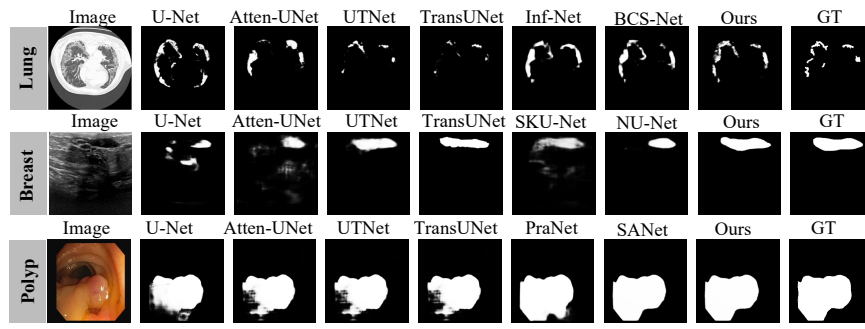


Fig. 5. Visual comparison of different medicine-general and medicine-specific methods.

5.4 Comparisons with State-of-the-art Methods

For the purpose of a comprehensive and unbiased comparison, we not only compare our approach with medicine-specific methods, but also with medicine-general methods, such as U-Net [24], Attention U-Net [19], UTNet [12], and TransUNet [7]. To be fair, we retrain these general-purpose methods using the same training data and settings as the proposed model. For different tasks, we use their common evaluation metrics. The results are shown in Table 1-4.

Table 1 shows performance comparison on the COVID-19 datasets. Compared to the second best method Inf-Net, our CHNet achieves an improvement of 1.7% in terms of mDice and the score has increased to 0.8. Table 2 shows that on the breast ultrasound dataset, our CHNet achieves the best performance in terms of all four metrics. Moreover, our model is more stable than the competitors, which can be verified by the standard deviations. In Table 3 and Table 4, our model consistently achieves the best segmentation performance on the polyp segmentation dataset ClinicDB and Kvasir and especially obtains significant improvement on mIoU metric.

Fig. 5 shows that the qualitative comparison with other methods. it can be seen that our results have greater advantages in terms of detection accuracy, object completeness, and contour sharpness across different image modalities.

Table 5. The FLOPs and parameters of different methods and their backbone.

Metrics	Backbone	U-Net	Atten-UNet	UTNet	TransUNet	Ours
FLOPs (GB)↓	~ 4.8	~ 17.5	~ 9.7	~ 27.1	~ 24.7	~ 7.1
Params (MB)↓	~ 8.7	~ 17.4	~ 19.2	~ 12.6	~ 34.0	~ 8.8

5.5 FLOPs and Parameters

In Table 5, we list FLOPs and parameters of different medicine-general methods and their backbone Res2Net50. It can be seen that the encoder and decoder of U-Net have used approximate amounts of parameter. The parameter number for Transformer-based TransUNet is even larger.

Compared to these U-Net-based models, our model achieves the best performance against both FLOPs and parameter amount metrics. The majority of parameters in the proposed method are derived from the backbone. The proposed model is efficient and its decoder is only about 0.15M.

5.6 Ablation Study

From Table 6, it can be seen that our classification head decoder has very strong feature expression ability. And only using 1×1 convolutional layer can achieve the best results, while using 3×3 convolutional layer performs slightly worse. The baseline adopts the FCN-32s [16] decoder.

With similar performance, 1×1 convolutional layers can save a lot of computational resources compared to 3×3 convolutional layers. Usually, applying a convolutional kernel with larger perception will cause the computational complexity to increase in the square mangificence.

Table 6. Ablation experiments on the COVID-19 Lung dataset.

Metrics	$mDice$	E_{ϕ}^{max}	F_{β}^{ω}	S_{α}
Baseline (FCN-32s)	0.666	0.601	0.496	0.799
CHNet (3×3 Conv)	0.796	0.761	0.653	0.843
CHNet (1×1 Conv)	0.800	0.765	0.655	0.846

6 Conclusion

In this paper, we propose a simple yet effective classification head network (CHNet) for medical image segmentation task, whose decoder is quite different from previous models. By conducting encoder-decoder decoupling experiments, we find that a decoder with complex structure will cause the feature extraction capability of the encoder to be under-utilized. Moreover, the decoupling experiments demonstrate that the feature expression capability of the classification head is similar to the well-designed decoders of other models, even though CHNet has very few parameters. More importantly, the feature extraction capability of the backbone is stronger for the end-to-end trained CHNet. Extensive experimental results on four datasets of three tasks demonstrate that the proposed model outperforms various state-of-the-art methods.

Acknowledgements. This work was supported by the National Natural Science Foundation of China # 62276046 and the Liaoning Natural Science Foundation # 2021-KF-12-10.

References

1. Covid-19 ct lung and infection segmentation dataset, <https://zenodo.org/record/3757476>, 2020, April
2. Covid-19 ct segmentation dataset, <https://medicalsegmentation.com/COVID19/>, 2020, April
3. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: International Conference on Learning Representations (2022), <https://openreview.net/forum?id=S1xSY2UZQT>
4. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilar-iño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics* **43**, 99–111 (2015)
5. Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., OBoyle, M., Comstock, C., Andre, M.: Breast mass segmentation in ultrasound with selective kernel u-net convolutional neural network. *Biomedical Signal Processing and Control* **61**, 102027 (2020)
6. Chen, G.P., Li, L., Dai, Y., Zhang, J.X.: Nu-net: An unpretentious nested u-net for breast tumor segmentation. *arXiv preprint arXiv:2209.07193* (2022)
7. 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)
8. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II* 19. pp. 424–432. Springer (2016)
9. Cong, R., Yang, H., Jiang, Q., Gao, W., Li, H., Wang, C., Zhao, Y., Kwong, S.: Bcs-net: Boundary, context, and semantic for automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Instrumentation and Measurement* **71**, 1–11 (2022)
10. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: *International conference on medical image computing and computer-assisted intervention*. pp. 263–273. Springer (2020)
11. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging* **39**(8), 2626–2637 (2020)
12. Gao, Y., Zhou, M., Metaxas, D.N.: Utinet: a hybrid transformer architecture 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 III* 24. pp. 61–71. Springer (2021)
13. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1055–1059. IEEE (2020)
14. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., Lange, T.d., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: *International Conference on Multimedia Modeling*. pp. 451–462. Springer (2020)

15. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: Resunet++: An advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM). pp. 225–2255. IEEE (2019)
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
17. Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L.: Y-net: joint segmentation and classification for diagnosis of breast biopsy images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11. pp. 893–901. Springer (2018)
18. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
19. 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)
20. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: CVPR. pp. 2160–2170 (2022)
21. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: CVPR. pp. 9413–9422 (2020)
22. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. IEEE TIP (2023)
23. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7479–7489 (2019)
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
25. Shin, S.Y., Lee, S., Yun, I.D., Kim, S.M., Lee, K.M.: Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. IEEE transactions on medical imaging **38**(3), 762–774 (2018)
26. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 699–708. Springer (2021)
27. Wei, J., Wang, S., Huang, Q.: F³net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12321–12328 (2020)
28. Wu, Z., Wang, L., Wang, W., Shi, T., Chen, C., Hao, A., Li, S.: Synthetic data supervised salient object detection. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5557–5565 (2022)
29. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: Datasetgan: Efficient labeled data factory with minimal human effort. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10145–10155 (2021)
30. Zhao, X., Jia, H., Pang, Y., Lv, L., Tian, F., Zhang, L., Sun, W., Lu, H.: M2snet: Multi-scale in multi-scale subtraction network for medical image segmentation. arXiv preprint arXiv:2303.10894 (2023)
31. Zhao, X., Pang, Y., Zhang, L., Lu, H.: Joint learning of salient object detection, depth estimation and contour extraction. IEEE TIP **31**, 7350–7362 (2022)

32. Zhao, X., Pang, Y., Zhang, L., Lu, H., Zhang, L.: Suppress and balance: A simple gated network for salient object detection. In: ECCV. pp. 35–51 (2020)
33. Zhao, X., Pang, Y., Zhang, L., Lu, H., Zhang, L.: Towards diverse binary segmentation via a simple yet general gated network. arXiv preprint arXiv:2303.10396 (2023)
34. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 120–130. Springer (2021)
35. 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, pp. 3–11. Springer (2018)