Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

μ-Net: Medical image segmentation using efficient and effective deep supervision

Di Yuan [a], Zhenghua Xu [a,*], Biao Tian [a], Hening Wang [a], Yuefu Zhan [b], Thomas Lukasiewicz [c,d]

[a] State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin, China
[b] Department of Radiology, Hainan Women and Children's Medical Center, Haikou, China
[c] Institute of Logic and Computation, TU Wien, Vienna, Austria
[d] Department of Computer Science, University of Oxford, Oxford, United Kingdom

ARTICLE INFO

Keywords:
Medical image segmentation
Deep supervised learning
Similarity principle of deep supervision
Tied-weight decoder

ABSTRACT

Although the existing deep supervised solutions have achieved some great successes in medical image segmentation, they have the following shortcomings; (i) semantic difference problem: since they are obtained by very different convolution or deconvolution processes, the intermediate masks and predictions in deep supervised baselines usually contain semantics with different depth, which thus hinders the models' learning capabilities; (ii) low learning efficiency problem: additional supervision signals will inevitably make the training of the models more time-consuming. Therefore, in this work, we first propose two deep supervised learning strategies, U-Net-Deep and U-Net-Auto, to overcome the semantic difference problem. Then, to resolve the low learning efficiency problem, upon the above two strategies, we further propose a new deep supervised segmentation model, called μ-Net, to achieve not only effective but also efficient deep supervised medical image segmentation by introducing a tied-weight decoder to generate pseudo-labels with more diverse information and also speed up the convergence in training. Finally, three different types of μ-Net-based deep supervision strategies are explored and a Similarity Principle of Deep Supervision is further derived to guide future research in deep supervised learning. Experimental studies on four public benchmark datasets show that μ-Net greatly outperforms all the state-of-the-art baselines, including the state-of-the-art deeply supervised segmentation models, in terms of both effectiveness and efficiency. Ablation studies sufficiently prove the soundness of the proposed Similarity Principle of Deep Supervision, the necessity and effectiveness of the tied-weight decoder, and using both the segmentation and reconstruction pseudo-labels for deep supervised learning.

1. Introduction

With the development of deep learning, deep-learning-based methods have been increasingly used in computer-aided medical diagnosis [1,2]. Medical image segmentation based on deep learning is one of the most important tasks [3], which aims to recognize and annotate the regions of interest (e.g., organs and lesions) with masks and/or outlines. U-Net [4] is one of the most popular deep-learning-based medical image segmentation techniques. It consists of a contracting path (encoder) to extract deep features from the input images, an almost symmetric expansive path (decoder) to achieve a precise localization, and skip connections to recover detailed image information lost during the down-sampling process [5]. Recent works have witnessed the application of U-Net in various medical image segmentation tasks, such as segmenting brain tumors [6,7], cardiac [8], liver [9], breast [10], and retinal vessels [11].

To improve the segmentation performance of U-Net, many existing works [12–14] propose to add additional supervision signals on the expansive path of U-Net to enforce direct and early supervision for both the intermediate layers and the output layer [15]. In [12,13], these works add a side output to each layer of the extensive path, i.e., directly up-sample the feature maps obtained by each layer to the target size by deconvolution, and then an additional intermediate loss is calculated in comparison with the ground truths; consequently, the model is trained based on a deep hybrid learning loss, obtained by summing up the intermediate losses on each expansive layers. In [14], it simply down-scales the ground truths to generate intermediate masks (i.e., applies max-pooling with fitting kernel size and stride to match the feature map's spatial extent) and then computes additional intermediate supervision losses by comparing the intermediate masks with the intermediate predictions at each layer of the expansive path.

[*] Corresponding author.
E-mail address: zhenghua.xu@hebut.edu.cn (Z. Xu).

https://doi.org/10.1016/j.compbiomed.2023.106963
Received 22 December 2022; Received in revised form 30 March 2023; Accepted 18 April 2023
Available online 23 April 2023

However, most of the existing deep supervised segmentation models encounter two same shortcomings as follows. (i) **Semantic difference problem:** Since the intermediate masks and the intermediate predictions used to estimate the intermediate losses in these methods are normally obtained by very different convolution or deconvolution processes, they usually contain semantics that belongs to different depths; consequently, directly comparing them to estimate intermediate losses for deep supervision may be arguable inappropriate, which will make the intermediate loss signals contain inaccurate guiding information, and thus not only limit the model's learning capability but also weaken the learning efficiency [16–18]. (ii) **Low learning efficiency problem:** Introducing additional intermediate learning signals can bring deep supervision with richer semantics into the learning process of segmentation models, and thus improve the model's segmentation accuracy to some extent; however, the additional supervision signals will be very likely to make the training of the models more time-consuming; the reason is as follows: by introducing additional intermediate learning signals, the time-cost of each training epoch of the deep supervised models will be significantly increased (due to the additional operations in the feed-forward process and the more complex objective function in the backpropagated optimization process), while the number of training epochs needed for the deep model to converge will also increase (in most cases) or remain similar (in even best case); so both factors work together will dramatically increase the total training time-cost in most scenarios.

Consequently, to overcome the above problems of the existing deep supervised segmentation models, in this work, we first propose two deep-supervised learning strategies of U-Net, *U-Net-Deep* and *U-Net-Auto*, to overcome the semantic difference problem and achieve deep supervision without any modification to the structure of U-Net. Then, we further propose a deep-supervised variant of U-Net, called *μ-Net*, which can resolve both the semantic difference and the low learning efficiency problems to achieve more accurate deep-supervised medical image segmentation with much lower training time-cost than the existing deep-supervised segmentation methods.

Specifically, we first propose **U-Net-Deep**, which is a simple but effective deep-supervised learning strategy of U-Net. U-Net-Deep requires no modification to the structure of U-Net, but simply adds an additional feed-forward step, which takes the segmentation mask (i.e., the ground truth) as the additional input of U-Net to generate intermediate feature maps at each layer of the expansive path as the intermediate pseudo-labels; consequently, U-Net-Deep additionally imports intermediate losses that measure differences between the intermediate pseudo-labels and the corresponding intermediate feature maps generated at the same layer of the expansive path using the medical image as inputs (called *intermediate outputs* for short) to achieve U-Net-based deep-supervised learning. Our experimental studies show that U-Net-Deep is not only simpler but also more effective than the existing deep-supervised variants of U-Net. We believe its effectiveness comes from its capability in resolving the semantic difference problem: the intermediate pseudo-labels of U-Net-Deep are obtained by convolution and deconvolution procedures that are identical to their corresponding intermediate outputs, making them contain the same depth of semantic information, so U-Net-Deep can estimate the intermediate losses more accurately than the baselines whose intermediate outputs and intermediate labels have semantic differences.

Despite resolving the semantic difference problem, we believe that the intermediate pseudo-labels generated by U-Net-Deep are still not perfect. This is because, although they are both images, the learning of medical images and segmentation masks in U-Net-Deep actually belongs to two different tasks: for medical images, the model aims to complete a segmentation task where the outputs are usually different to the inputs; but when using segmentation masks as inputs of U-Net, the model aims to generate outputs that are the same as inputs, which is actually a reconstruction task [19]. Consequently, we further propose another deep-supervised learning strategy of U-Net, **U-Net-Auto**, which uses an independent autoencoder (whose structure is almost the same as that of U-Net, except for the lack of skip connections) to accomplish the reconstructive learning process of segmentation masks and obtain intermediate pseudo-labels (at the decoder) for deep supervision. Experimental studies show that U-Net-Auto consistently outperforms U-Net-Deep, we believe this is because of the following two reasons: (i) the autoencoder in U-Net-Auto have almost the same structure as U-Net, making the generated intermediate pseudo-labels still have the same depth of semantics as the intermediate outputs; and (ii) use task-specific models U-Net and autoencoder to process the input images and labels respectively can help the obtained intermediate pseudo-labels and intermediate outputs keep diverse features and accommodate with their own tasks.

Although U-Net-Deep and U-Net-Auto properly resolve the semantic difference problem, the low learning efficiency problem has not been fully addressed yet. Our experimental studies show that by resolving the semantic difference problem, U-Net-Deep and U-Net-Auto not only can enhance and increase the models' feature learning capability and achieve much better segmentation performances, but the numbers of training epochs needed for their model convergences are also lower than those of the existing deep supervised methods; however, since U-Net-Deep and U-Net-Auto require to take two inputs, their average time-cost for each training epoch is higher; consequently, when the increase of the latter overwhelms or is similar to the decrease of the former, the total training time-cost of U-Net-Deep may be similar or even higher than the existing deep supervised methods. To fully resolve the low-efficiency problem, we further propose an efficient and accurate deep supervised segmentation model, **μ-Net**, to dramatically further decrease the numbers of convergence epochs and ensure that the decrease of epochs is always much higher than the increase of time-cost per epochs; consequently, μ-Net will have much higher efficiency than the existing deep supervised models, and thus fully resolve the low-efficiency problem.

Generally, μ-Net proposes to share the same encoder for both U-Net and autoencoders, but retain the different decoders for them, where the decoder of the autoencoder is a tied-weight decoder [20], i.e., the weight matrices in the decoder are the transposes of those in the encoder. Consequently, μ-Net has the following two advantages: (i) Sharing the same encoder for both U-Net and autoencoder makes the generated intermediate masks not only have the same depth of semantics with the generated intermediate outputs, but they also rely on the same group of learned features (i.e., the same encoder), while retaining different learning processes in decoders according to different tasks. Consequently, μ-Net not only resolves the semantic difference problem, but its intermediate masks and intermediate outputs also own consistency in feature learning and diversity in image restoration and prediction, which thus makes its deep supervisions more accurate and effective. (ii) The usage of tied-weight makes the decoder of autoencoder actually share the same weights with the encoder, which thus greatly alleviates the vanishing gradient problem and speeds up the model's convergence; therefore, μ-Net is capable to resolve the low learning efficiency problem and achieve efficient deep supervision. Specifically, given an input image and its corresponding label as the inputs, the intermediate outputs of input images are obtained at each layer of the U-Net's decoder; however, we can obtain two groups of intermediate pseudo-labels: one group is generated at the segmentation decoder of U-Net (thus called *segmentation pseudo-labels*), while the other is generated at the reconstruction decoder of autoencoder (thus called *reconstruction pseudo-labels*). Consequently, there exist three kinds of deep supervision strategies for μ-Net: (i) using solely segmentation pseudo-labels for deep supervision, denoted ***μ-Net-Seg***; (ii) using solely reconstruction pseudo-labels for deep supervision, denoted ***μ-Net-Rec***; and (iii) using both for deep supervision, denoted ***μ-Net-Hyb***.

Our experimental studies show: (i) μ-Net-Hyb achieves the best performances, which proves that the segmentation and the reconstruction pseudo-labels are beneficial for deep supervised learning, they can complement each other to achieve superior performances. (ii) μ-Net-Rec generally outperforms μ-Net-Seg; we believe this is due to the same reason why U-Net-Auto achieves better performances than U-Net-Deep; consequently, according to these observations, we thus summarize a **Similarity Principle of Deep Supervision** to measure the quality of the intermediate learning signals used in deep supervised models. Specifically, the Similarity Principle of Deep Supervision can be described as follows. To obtain good intermediate learning signals, the intermediate masks (i.e., pseudo-labels) and their corresponding intermediate outputs (i.e., side outputs) should be generated using similar but NOT identical convolution and deconvolution procedures (i.e., the same number of operations but using different convolution and/or deconvolution layers); we believe this will ensure the intermediate masks and intermediate outputs contain semantics at the same depth (thus avoid the semantic difference problem) and will also help them learn more diverse and task-specific features than directly using identical convolution and deconvolution procedures to obtain them. Finally, please note that, to keep it simple, we use U-Net as the backbone in our methodology, however, μ-Net is a highly adaptable and scalable solution that can be applied to all U-Net-based advanced segmentation models, e.g., U-Net++ [17] and Attention U-Net [21], to achieve better segmentation performances.

The main contributions of this paper are briefly as follows:

- We identify the semantic difference and low learning efficiency problems of the existing deep supervised segmentation methods.
- We first propose two deep supervised learning strategies, U-Net-Deep and U-Net-Auto, to overcome the semantic difference problem. Then, to resolve the low learning efficiency problem, upon the above two strategies, we further propose a new deep supervised segmentation model, called μ-Net, to achieve not only effective but also efficient deep supervised medical image segmentation by introducing a tied-weight decoder to generate pseudo-labels with more diverse information and also speed up the convergence in training. Finally, three different types of μ-Net-based deep supervision strategies are explored and a Similarity Principle of Deep Supervision is further derived to guide future research in deep supervised learning.
- Experimental studies on four public benchmark datasets show that μ-Net greatly outperforms all the state-of-the-art baselines, including the state-of-the-art deeply supervised segmentation models, in terms of both effectiveness and efficiency. Ablation studies sufficiently prove the soundness of the proposed Similarity Principle of Deep Supervision, the necessity and effectiveness of the tied-weight decoder, and using both the segmentation and reconstruction pseudo-labels for deep supervised learning.

The rest of this paper is organized as follows. Literature reviews are included in Section 2, while the details of μ-Net are introduced in Section 3. Section 4 then presents the experimental studies and evaluates the supremacy of μ-Net and the soundness of the proposed Similarity Principle of Deep Supervision. Finally, we conclude the paper and discuss potential future works in Section 5.

2. Related work

Medical Image Segmentation. With the development of deep learning, more and more deep models are successfully applied to medical image segmentation [22]. FCN is the first end-to-end image segmentation model using convolutional neural networks [23]; FCN-based medical image segmentation is mainly achieved by first using convolution and pooling operations for feature learning and then applying a transpose convolutional up-sampling-based skip architecture for pixel-level classifications [24,25]. To obtain a more refined segmentation, U-Net is further proposed to upgrade FCN to a structure with symmetrical contracting (down-sampling) and expansive (up-sampling) paths, and skip connections are also used in U-Net to concatenate the deep and coarse features in the expansive path with the shallow and fine features in the contracting path for more accurate and detailed segmentation [4]. U-Net is arguably the most widely adopted deep model for medical image segmentation; recent works witness the application of U-Net in various segmentation tasks, such as brain tumor [6,7], cardiac [8], liver [9,13], and retinal vessel [11,26] segmentation.

To further improve the segmentation performance, many advanced variants of U-Net are proposed in recent works [18,27,28]. Attention U-Net is proposed in [21] to segment the pancreas on CT images, where attention gates [29] are integrated into the expansive path to suppress the response of irrelevant background information and enhance the sensitivity of the pancreas features via assigning different weights. ResUNet++ [30] is proposed for colonoscopic image segmentation, where three improving techniques (i.e., residual connections, attention mechanism, and atrous spatial pyramidal pooling) are incorporated into U-Net to enhance the segmentation performance. To show the superior performance of μ-Net in medical image segmentation, these state-of-the-art medical image segmentation techniques, i.e., FCN [23], U-Net [4], Attention U-Net [21], and ResUNet++ [30], are selected as the baselines in our work.

Deep Supervised Segmentation. To improve the segmentation performance of U-Net, deep supervision was adopted in many existing works [31–34], which add additional supervision signals on the hidden layers of U-Net to enforce direct and early supervision for both the hidden layers and the output layer [15]. An edge-aware mechanism [35] was proposed to segment retinal vessels on fundus images, where the fused outputs of four side-output layers from both the contracting and the expansive path are used to compare with the ground truths to calculate auxiliary losses to help the networks converge. U-Net++ [17] introduced a deep supervision technique to achieve a better performance based on multiple medical image segmentation tasks, adding a supervisory signal after each nested dense convolution of the first layer to supervise the output of U-Net for each branch to enhance the segmentation performance. A 3D DSN [36] was proposed to segment CT and MR medical images, up-scaling hidden-level features using additional deconvolutional layers to counteract adverse effects of unstable gradient changes. Besides, to solve the problem of information loss during forward propagation, a deeply supervised nonlinear aggregation model [34] was proposed for salient object detection, where side-output features are from the expansive path, and are aggregated in a nonlinear way to calculate the losses relative to the ground truths. As the main difference to these works, however, our deep supervised signals are added in each expansive hidden layer of U-Net for medical image segmentation.

Similar to our work, there also exists research achieving deep supervision by adding additional supervision signals on each expansive hidden layer of U-Net. UNet3+ [13] generates intermediate outputs by bilinear up-sampling to learn hierarchical representations from full-scale aggregated feature maps. M-Net [12] uses the side-output to produce a companion local prediction map for different scale layers and uses a multi-label loss function to calculate the additional supervision signals. Li et al. [37] use the auxiliary outputs that are added before each upper sampling to supervise the model deeply and mask gradient propagation better. Liu et al. [38] compare the ground truths with the merged outputs of the last layer and of the intermediate layer after up-sampling. In addition to the above methods where the feature maps are generated by each expansive hidden layer after deconvolution and up-sampling, and then directly compared with the ground truths to obtain additional supervision signals, an alternative is to downsample the ground truths to a size similar to the feature maps generated by the corresponding expansive hidden layer as the pseudo-labels. The former is called *upsampled deep supervision*, while the latter is called *downsampled deep supervision*. Reiss et al. [14] propose a new supervision mechanism, which down-samples the ground truths to match the feature map's spatial extent through a suitable max-pooling layer.

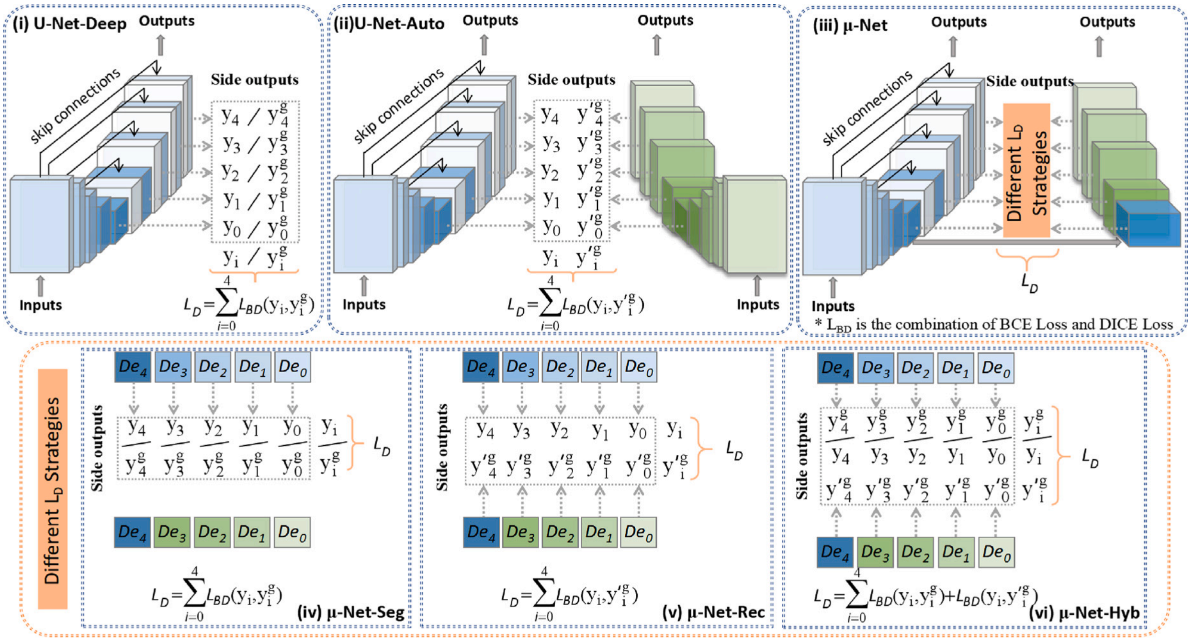

(i) U-Net-Deep
Outputs
skip connections
Side outputs
y_4 / y_4^g
y_3 / y_3^g
y_2 / y_2^g
y_1 / y_1^g
y_0 / y_0^g
y_i / y_i^g
Inputs
$L_D=\sum_{i=0}^{4} L_{BD}(y_i, y_i^g)$
(ii)U-Net-Auto
Outputs
Outputs
skip connections
Side outputs
y_4 y'^g_4
y_3 y'^g_3
y_2 y'^g_2
y_1 y'^g_1
y_0 y'^g_0
y_i y'^g_i
Inputs
Inputs
$L_D=\sum_{i=0}^{4} L_{BD}(y_i, y'^g_i)$
(iii) μ-Net
Outputs
Outputs
skip connections
Side outputs
Different L_D Strategies
L_D
Inputs
* L_{BD} is the combination of BCE Loss and DICE Loss
Different L_D Strategies
Side outputs
De_4 De_3 De_2 De_1 De_0
$\frac{y_4}{y_4^g}$ $\frac{y_3}{y_3^g}$ $\frac{y_2}{y_2^g}$ $\frac{y_1}{y_1^g}$ $\frac{y_0}{y_0^g}$ $\frac{y_i}{y_i^g}$ L_D
De_4 De_3 De_2 De_1 De_0
$L_D=\sum_{i=0}^{4} L_{BD}(y_i, y_i^g)$
(iv) μ-Net-Seg
Side outputs
De_4 De_3 De_2 De_1 De_0
y_4 y_3 y_2 y_1 y_0 y_i
y'^g_4 y'^g_3 y'^g_2 y'^g_1 y'^g_0 y'^g_i L_D
De_4 De_3 De_2 De_1 De_0
$L_D=\sum_{i=0}^{4} L_{BD}(y_i, y'^g_i)$
(v) μ-Net-Rec
Side outputs
De_4 De_3 De_2 De_1 De_0
y_4^g y_3^g y_2^g y_1^g y_0^g y_i^g
y_4 y_3 y_2 y_1 y_0 y_i L_D
y'^g_4 y'^g_3 y'^g_2 y'^g_1 y'^g_0 y'^g_i
De_4 De_3 De_2 De_1 De_0
$L_D=\sum_{i=0}^{4} L_{BD}(y_i, y_i^g)+L_{BD}(y_i, y'^g_i)$
(vi) μ-Net-Hyb

Fig. 1. Overall structures of the proposed U-Net-Deep, U-Net-Auto, and μ-Net, where three different deep supervision strategies for μ-Net are also visualized. Here, L_{BD} represents the combination of BCE and Dice Losses.

However, we found that both the existing upsampled and downsampled deep supervision methods will encounter the problem of semantic difference, so in this work, we propose a deep supervision semantic similarity principle, and based on this principle, we propose that μ-Net makes additional supervision on each expansive hidden layer calculated based on intermediate predictions and pseudo-labels of similar semantics. To show the superior performance of our Similarity Principle of Deep Supervision in segmentation, we conduct experiments comparing our method with the main existing deep supervisions: upsampled deep supervision, M-Net [12], downsampled deep supervision, and MLDS-Net [14].

Integrating Autoencoders with U-Net. Similar to our work, there are other works that integrate an autoencoder with U-Net to achieve a more accurate and efficient segmentation. Myronenko et al. [39] add a variational autoencoder branch to the encoder endpoint to reconstruct the original image to regularize the tied decoder and impose additional constraints on its layers. A hierarchical probabilistic U-Net [40] is a U-Net with a conditional variational decoder to achieve a high-fidelity sampling and reconstruction while providing flexibility. In [41], a variational decoder can correct many topological incoherences by learning a rich but compact latent space. However, our work is different from these works for the following reason: most of these methods use the autoencoder to improve the speed and accuracy of feature learning by the introduced reconstructive loss. While in μ-Net, in addition to the above purpose, we also use the autoencoder to generate pseudo-labels that are semantically similar to but not completely consistent with the immediate predictions on the expansive path of U-Net, so we use a tie-weight instead of a variational autoencoder. This is to ensure that the expansive path is structurally similar to the shared extracting path and the expansive path of U-Net, to ensure that the pseudo-labels and intermediate predictions have a similar semantics.

3. Methodologies

U-Net [4] and its variants use the contracting path to capture semantic features and the expansive path to restore the location information and introduce skip connections to recover the lost information. However, these models may not be sufficient to fully recover the lost information and inevitably result in inaccurate segmentation for small objects and objects with complex boundary details [42]. Intuitively, in the contracting path of U-Net, the features of these important small objects and edges become less and less visible or even disappear, leading to an inaccurate segmentation for small objects and objects with complex boundary details [43]. These minor errors may be tolerable in natural image segmentation, but are unacceptable in medical image segmentation, because they may have fatal consequences in clinical practice (e.g., when this model is used to delineate the object area of a tumor lesion in radiotherapy, even a few tumor cells missed may cause the failure of radiotherapy and cancer recurrence). Therefore, the need for a more accurate deep model for medical image segmentation is compelling. The commonly used solution is to use deep supervision to provide semantic information to the intermediate layers, and to improve the segmentation performance by directly and early constraining the weight update process of the intermediate layers and the output layer. Most of these methods use the ground truths through convolutional layers or pooling operations to obtain pseudo-labels, and then compare these pseudo-labels with the intermediate outputs obtained through multiple layers of down-sampling and up-sampling. However, the pseudo-labels and intermediate predictions in these methods have large semantic differences due to the different ways of generating them. Besides, deep intermediate learning with additional supervision signals will inevitably mask the training of the models more time-consuming.

Hence, in this paper, we propose a series of models to solve the above problems; see Fig. 1. Compared with the existing deep supervision, our models mainly obtain the pseudo-labels by the ground truths through down-sampling and up-sampling similar to the intermediate outputs, and then compare the pseudo-labels with the intermediate predictions to constrain the weight filtering process of the intermediate layers. Intuitively, the main idea is that the pseudo-labels and intermediate predictions can make up the semantic difference between the two by going through similar, rather than identical, convolutional processes. Specifically, we first propose U-Net-Deep (pseudo-labels and intermediate outputs are generated through a fully consistent convolution process) and U-Net-Auto (pseudo-labels are generated through an autoencoder similar to the segmentation model) to overcome the semantic difference problem. In addition, we further propose a more effective and efficient segmentation model, μ-Net, which not only further expands the diversity of details of pseudo-labels but also improves the convergence performance via the tied-weight decoder. According to the different pseudo-labels used, μ-Net can be refined into three training strategies.

3.1. U-Net-Deep

In this section, we propose U-Net-Deep based on U-Net to overcome the semantic difference problem, shown in Fig. 1(i). More details of the proposed U-Net-Deep and the corresponding hybrid loss function are presented in the following subsections.

In detail, as shown in Fig. 1(i), similarly to U-Net, U-Net-Deep feeds the original images into U-Net and then outputs the segmentation results; but, differently from U-Net, U-Net-Deep also feeds the corresponding ground truths into U-Net as the additional inputs. Then, the intermediate outputs generated by the original images and ground truths in the expansive path are used as intermediate segmentation outputs and intermediate segmentation masks, respectively, and the loss between the two is called the intermediate loss function. Finally, U-Net-Deep is learned by comprehensive supervision, that is, mask-guided supervision signals are added to each layer of the expansive layer.

The advantage of U-Net-Deep is that the ground truths are directly fed into the segmentation model to generate supervision signals in each layer of the expansive path. On the one hand, the existing deep supervision methods (i.e., U-Net++) add some supervision signals to the convolution block of the last layer. These methods do not guarantee that the features of small objects or segmentation edges are still completely preserved in the middle layer. On the other hand, the existing deep supervision methods add supervision signals to the outputs of the intermediate layer after a series of convolution operations. These methods are likely to cause a further loss of important features. Our method is to directly feed the ground truths into the model, and use the masks to guide the comprehensive intermediate supervision mechanism by adding supervision signals to each layer of the expansive path, which can more effectively improve the segmentation performance of small objects and segmentation edges. Our methods are described in detail as follows.

Our proposed intermediate supervision mechanisms all adopt a set of 2D grayscale medical images as the inputs for the segmentation model to reduce the model's calculation efforts. To obtain the 2D grayscale medical images, we first transform all the original datasets into 2D medical images according to the transverse section in a slice-by-slice manner [44]. However, medical images and segmentation masks in BraTS are RGB images. Therefore, to confirm that the inputs are 2D grayscale images, we then use PyTorch to convert all inputs into grayscale images before all 2D medical images are fed into models. Note that to make the difference between the segmentation outputs and the segmentation backgrounds more visible, we convert all the output grayscale images into RGB images after the last layer of the model.

After that, like in the original U-Net, we input the original medical images (denoted as x) into the segmentation model to produce the segmentation outputs (denoted as y) and calculate the loss L_{image} of them. Besides, to add some additional intermediate supervision signals, we also preserve the intermediate segmentation outputs, denoted as y_i (where $i \in \{0,1,2,3,4\}$). However, the intermediate segmentation outputs are segmented from the original images through a series of convolution and non-linearities operations [21]. So, their sizes are different from those of the corresponding original segmentation masks. Thus, one of the problems that we face is how to get the corresponding ground truths of the intermediate segmentation outputs.

It occurred to us that if we get the intermediate outputs of the corresponding segmentation masks, this problem will be solved. Therefore, we also input the corresponding segmentation masks G into the model as additional inputs to generate the intermediate segmentation outputs y_i^g and the final outputs y^g of the segmentation masks. Besides, y_i^g are regarded as the corresponding ground truths of y_i. It is worth noting that these ground truths are transformed into grayscale images before they are as additional inputs fed into the network. Finally, the segmentation loss between y and G is regarded as L_{image}, and L_{mask} is the segmentation loss between y^g and G, and L_{seg} is the segmentation loss between y and y^g. Moreover, we calculate the additional intermediate supervision losses L_i of the deepest layer and the expansive path of the model according to y_i and y_i^g. Briefly, based on U-Net, our U-Net-Deep adds some intermediate supervision signals (i.e., L_i) and a final layer loss (i.e., L_{seg} and L_{mask}) for small objects in medical images to improve the segmentation accuracy of the model for small objects. In detail, these intermediate supervision losses are as follows.

Formally, given the original medical images X, the corresponding segmentation masks G, and their outputs in the last layer of the model denoted as y and y^g, the above loss functions are defined as follows:

$$L_{image} = L_{BD}(y, G), \tag{1}$$

$$L_{mask} = L_{BD}(y^g, G), \tag{2}$$

$$L_{seg} = L_{BD}(y, y^g), \tag{3}$$

where L_{BD} is the combination of binary cross-entropy loss [45] (BCE loss) and Dice loss [46] (also known as F1 score), which is denoted as follows:

$$L_{BD}(y, y^g) = -\frac{1}{N}\sum_{j}^{N}\left(\frac{1}{2}\cdot y_{[j]}\cdot \log y^g_{[j]} + \frac{2\cdot y_{[j]}\cdot y^g_{[j]}}{y_{[j]} + y^g_{[j]}}\right), \tag{4}$$

where $y_{[j]}$ and $y^g_{[j]}$ denote the predicted probabilities and the ground truths of the jth image, respectively, and N indicates the batch size. Then, the intermediate outputs of X and G of the segmentation model are denoted as y_i and y_i^g, respectively, and the deep loss function in the intermediate layers is denoted as follows:

$$L_D = L_M(y_i, y_i^g) = \frac{1}{N}\sum_{j}^{N}(y_{i[j]},\ y^g_{i[j]})^2, \tag{5}$$

where L_M is the mean squared error (MSE) loss [47], $i \in \{0,1,2,3,4\}$. Finally, our hybrid loss function (L_h^1) is the sum of L_{image}, L_{mask}, L_D, and L_{seg}. In this way, the loss function not only considers the segmentation outputs in the deepest layer of the model like in the original U-Net, but also adds multiple intermediate supervision signals in the intermediate layers of the model. Therefore, the model's learning ability for small objects and the segmentation details in medical images can be enhanced. L_h^1 is defined as follows:

$$L_h^1 = \alpha L_{image} + \beta L_{mask} + \lambda \sum_{i=0}^{4} \omega_i L_D + \gamma L_{seg}, \tag{6}$$

where α, β, λ, ω_i, and γ are the weights of the loss functions, which are independent parameters that can be adjusted as required.

3.2. U-Net-Auto

However, U-Net is mainly engaged in segmentation, and the extracted feature is also the position information for precise positioning, and the input and output of U-Net are different images. Our intermediate monitoring mechanism requires both the original mask input and output, which is more similar to the reconstruction process of an autoencoder (AE) [48]. Therefore, we combine an AE with U-Net and propose U-Net-Auto. The structure is shown in Fig. 1(ii). The main difference to U-Net-Deep is as follows. U-Net-Deep directly inputs corresponding segmentation masks into the model, and the intermediate output generated by them is taken as corresponding ground truths, while U-Net-Auto inputs segmentation masks into another autoencoder, and takes the intermediate output generated in the autoencoder as corresponding ground truths. U-Net-Auto and its loss function are as follows.

Similarly to U-Net-Deep, the original image X and the corresponding segmentation masks G are input into the AE to generate reconstructed outputs x and y'^g, and intermediate outputs y_i and $y_i'^g$, respectively. $y_i'^g$ and y'^g are also seen as the corresponding ground truths of y_i and y, respectively. Intuitively, we can think that the middle ground truths can be obtained by minimizing the reconstruction errors. In addition, we also input original medical images into the AE to enhance the effect of reconstruction of the AE. Finally, the loss function L'_{image} between y and G in U-Net, the loss function L'_{rem} between y'^g and G in the AE, the loss function L'_{rei} between x (the outputs of original images) and X in the AE, the loss function L'_{seg} between y and y', and the intermediate loss function L'_D between y_i and $y_i'^g$ are superimposed together as the hybrid loss function of U-Net-Auto:

$$L_h^2 = \alpha L'_{image} + aL'_{rem} + bL'_{rei} + \gamma L'_{seg} + \lambda \sum_{i=0}^{4} \omega_i L'_D, \tag{7}$$

$$L'_{rem} = L_B(y'^g, G), \tag{8}$$

$$L'_{rei} = L_B(x, X), \tag{9}$$

$$L_B(\widetilde{x}, g) = -\frac{1}{N}\sum_{j}^{N}[\widetilde{x}_{[j]} \log g_{[j]} + (1 - \widetilde{x}_{[j]}) \log(1 - g_{[j]})], \tag{10}$$

where L_B is the BCE loss, $\widetilde{x}$ and g denote the reconstructed probabilities and the ground truths of the jth image, respectively, and N indicates the batch size. α, a, b, γ, λ, and ω_i are the weights of the loss functions, which are independent parameters that can be adjusted as required.

3.3. μ-Net

Although U-Net-Deep and U-Net-Auto properly solve the semantic difference problem, the low learning efficiency problem has not been completely solved. Specifically, by solving the semantic difference problem, U-Net-Deep and U-Net-Auto can not only enhance the model's feature learning ability and achieve better segmentation performance, but also the number of training epochs required for achieving convergence is lower than the existing deep supervision methods. However, because U-Net-Deep and U-Net-Auto require two inputs, they have a higher average time-cost for each training epoch. Thus, when the increase of the latter overwhelms or is similar to the decrease of the former, their total training time-cost may be similar to or even higher than existing deep supervision methods. Besides, U-Net-Deep and U-Net-Auto have two different encoders, which may lead to different extracted features (for example, one encoder may extract horizontal features, while another one may extract vertical features), which may lead to the inconsistent output of the decoders. The resulting intermediate segmentation mask is not the most ideal [49]. To fully solve the low-efficiency problem and further improve the segmentation performance, we propose an efficient and accurate deep supervised segmentation model, μ-Net, to further significantly reduce the number of convergence epochs and ensure that the epoch reduction is always far higher than the increase of each epoch time-cost. Consequently, μ-Net completely solves the low-efficiency problem and further improves the segmentation performance. μ-Net combines the encoders of U-Net and AE to form a tied-weight decoder whose structure is shown in Fig. 1(iii).

Intuitively, for the following reason, adding the tied-weight decoder can greatly accelerate the model's learning progress and reduce training runs needed for model convergence. First, autoencoders can extract useful features continuously during backpropagation and filter the useless information [48]. Then, in the intermediate supervision mechanisms, the learning signals of the segmented objects become very weak when they are backpropagated to the first few layers, so learning the first few weight matrices (e.g., W_1 and W_2) is very slow, which is the vanishing gradient problem [20]. In μ-Net, we use a tied-weights decoder, which means that the weight matrices in the expansive path are the transposes of those in the contracting path. Therefore, the reconstruction-error-based learning signal will be used to first update W_1^T, then backpropagate to update the second layer W_2^T, the third layer W_3^T, and so on. As updating W_i^T (where $i \in \{0, 1, 2, 3, 4\}$) is equivalent to updating W_i, it remedies the vanishing gradient problem.

Fig. 1(iii) shows the overall process of the proposed μ-Net. μ-Net is similar to U-Net-Deep and U-Net-Auto, but adds four convolution blocks to form an additional decoder and uses one contracting path, so, by taking the first five blocks as encoder, we convert each convolution block to an autoencoder with *tied weights*. That is, the weight matrices in the additional decoder are the transposes of those in the original encoder. It is important to note that μ-Net is different from U-Net-Auto in the AE branch. The AE branch in U-Net-Auto is designed to ensure that the same features are extracted from the U-Net and AE in the contracting path to improve the segmentation performance. Whereas the AE branch of μ-Net is to improve the model's training efficiency while improving the segmentation performance. According to the intermediate outputs generated by the ground truths in different expansive paths, μ-Net can be divided into three substructures: μ-Net-Seg, μ-Net-Rec, and μ-Net-Hyb, as shown in Fig. 1(iv)–(vi).

In μ-Net-Seg, the original images X are first fed into the model to generate the segmented outputs y in U-Net and the reconstructed outputs y^x in AE. Meanwhile, the intermediate outputs y_i and the intermediate reconstruction y_i^x are generated in the two branches. Then, the corresponding ground truths G are fed into the model, generating the outputs y^g and intermediate outputs y_i^g in the branch of U-Net. Similarly, y_i^g are the ground truths of y_i. Finally, the loss function L^1_{image} between y and G in U-Net, the loss function L^1_{rei} between y^x and X in AE, the loss function L^1_{mask} between y^g and G in U-Net, and the intermediate loss function L^1_D between y_i and y_i^g, and L^1_{seg} between y and y^g are superimposed together as the hybrid loss function of μ-Net-Seg:

$$L_h^{31} = \alpha L^1_{image} + \beta L^1_{mask} + bL^1_{rei} + \gamma L^1_{seg} + \lambda \sum_{i=0}^{4} \omega_i L^1_D, \tag{11}$$

where α, λ, b, γ, λ, and ω_i are the weights of the loss functions, which are independent parameters that can be adjusted as required.

The structure of μ-Net-Rec is similar to μ-Net-Seg, the only difference is that the intermediate outputs $y_i'^g$ and y'^g are generated by inputting the ground truths G into the expansive path of AE. The hybrid loss function is the sum of the loss function L^2_{image} between the segmentation outputs y and G in U-Net, the loss function L^2_{rem} between the reconstruction outputs y^x and G in AE, the loss function L^2_{rei} between y^x and X, the intermediate loss L^2_D between the intermediate outputs y_i in U-Net and the intermediate outputs $y_i'^g$ in AE, and the intermediate loss L^2_{seg} between the outputs y in U-Net and the outputs r in AE:

$$L_h^{32} = \alpha L^2_{image} + aL^2_{rem} + bL^2_{rei} + \xi L^2_{seg} + \eta \sum_{i=0}^{4} \sigma_i L^2_D, \tag{12}$$

where α, a, b, ξ, η, and σ_i are the weights of the loss functions, which are independent parameters that can be adjusted as required.

μ-Net-Hyb can be regarded as a comprehensive structure of μ-Net-Seg and μ-Net-Rec. It not only includes the intermediate supervision signals generated by the expansive paths of U-Net in μ-Net-Seg, but also includes the intermediate supervision signals generated by the expansive path of AE in μ-Net-Rec. The hybrid loss function is composed of (i) the loss function L^3_{image} between the segmentation outputs y and G, the loss function $L^3 mask$ between y^g and G in U-Net; (ii) the loss function L^3_{rem} between the reconstruction outputs y^x and G, the loss function L^3_{rei} between y^x and X in AE; (iii) the intermediate loss function L^3_D between the intermediate outputs y_i and y_i^g in U-Net, and the intermediate loss function $L^{3'}_D$ between the intermediate outputs y_i in U-Net and $y_i'^g$ in AE; and (iv) the loss function L^3_{seg} between y and y^g in U-Net, and the

Table 1
Datasets information.

Datasets	Images	Input size	Modality	Challenge	Source
BraTS [50]	24,864	240 × 240	T1ce	Complex and heterogeneously-located objects	MICCAI
Cardiac [51]	1350	320 × 320	MRI	Small training dataset with large variability	King's College London
Spleen [51]	1050	512 × 512	CT	Large ranging foreground size	Memorial Sloan Kettering Cancer Center
Liver [51]	19,160	512 × 512	CT	Label unbalance with a large and small target	Several clinical sites

loss function $L^{3'}_{seg}$ between y in U-Net and y'^{g} in AE:

$$L_h^{33} = \alpha L^3_{image} + \beta L^3_{mask} + a L^3_{rem} + b L^3_{rei} + \gamma L^3_{seg} + \xi L^{3'}_{seg} + \lambda \sum_{i=0}^{4} \omega_i L^3_D + \eta \sum_{i=0}^{4} \sigma_i L^{3'}_D. \quad (13)$$

4. Experimental studies

Extensive experiments have been conducted to evaluate our proposed μ-Net. In this section, we first introduce the information of datasets, baselines, experimental settings, and evaluation metrics in Sections 4.1–4.3. Then, to prove the effectiveness of our method, we have conducted extensive experimental studies to compare the performance of μ-Net with four state-of-art baselines: FCN [23], U-Net [4], Attention U-Net [21], and ResUNet++ [30]. After that, to validate the effectiveness and necessity of the Similarity Principle of Deep Supervision and tied-weight decoder in μ-Net, ablation studies are further conducted. Furthermore, to show that our proposed Similarity Principle of Deep Supervision is more suitable for small object segmentation on medical images than the state-of-the-art deep supervision mechanisms, additional experiments are conducted to compare the performance of using the Similarity Principle of Deep Supervision with that of using two deep supervision mechanisms. Similarly, to show that the tied-weight decoder is more effective for small object segmentation on medical images than the common VAE decoder, we have conducted supplementary experiments to compare the Similarity-Principle-of-Deep-Supervision-based tied-weight decoder with the similarity-principle-of-deep-supervision-based VAE decoder. Finally, to obtain the optimal hyperparameter settings in the proposed Similarity Principle of Deep Supervision and tied-weight decoder and to achieve the best performances for μ-Net, grid search has been applied to evaluate the effect of varying loss weights on the performances of μ-Net.

4.1. Datasets and preprocessing

The empirical studies over four real datasets confirm that our models beat other baseline models. As shown in Table 1, we use four medical imaging datasets for model evaluation, covering lesions/organs from different medical imaging modalities. These datasets contain the characteristics of small datasets and small objects and are more representative of the characteristics of current medical images. Moreover, these datasets have a common feature: the segmentation objects are smaller than the background image and contain more segmentation details (e.g., the segmentation edges).

BraTS (brain tumor segmentation) [50] uses the HGG in the BraTS 2019 training set, because the test set has no labels, and the tumor in the HGG is relatively obvious. BraTS has 259 cases, each case contains 155 slices of 240 ∗ 240 MR images. Each case has four modes (T1, T2, Flair, T1ce), we choose the T1ce mode to segment the whole tumor (WT), enhance tumor (ET), and tumor core (TC). The difficulty of segmentation on BraTS lies in the complex and heterogeneously-located objects. The segmentation task on BraTS is challenging because it not only contains 3 kinds of segmenting objects, i.e., whole tumor region, tumor core region, and tumor enhancement region, but also most of the tumor enhancement regions are very small, taking up only hundreds of pixels in the images (see the white small spots in the first two rows of Fig. 2 as examples).

Cardiac [51] is a public CT dataset to automatically segment the heart, which contains 20 cases. As the scanning mechanism is different, each case has 320 ∗ 320 CT images with a range of 90 to 130 slices. The difficulty of its segmentation is that the dataset is small and the segmentation objects change greatly.

Spleen [51] is a public CT dataset to automatically segment the spleen, which contains 41 cases. As the scanning mechanism is different, each case has 512 ∗ 512 CT images with a number of slices ranging from 31 to 168. The segmentation challenge of this dataset is that the segmentation objects vary greatly.

Liver [51] is a public liver CT dataset which contains 131 cases. Each case contains some 512 ∗ 512 CT images (i.e., slices), and the number of slices ranges from 29 to 299. Similar to BraTS, the segmentation task on this dataset is twofold: aiming to segment not only livers but also the corresponding liver tumors, where the sizes of liver tumors are relatively small, i.e., taking up 3689 pixels in average (i.e., only about 1.4% of the whole image).

To train and evaluate the networks, all the above four medical image datasets are preprocessed as follows. First, for all datasets, we transform these $3D$ images into $2D$ images according to the transverse section in a slice-by-slice manner [44]. Then, we normalize all input images to have zero mean and unit std. After that, for the BraTS dataset, to remove a part of the images without the segmentation objects, we removed the 30 slices before and after, that is, only the 30th to the 125th slices are selected in our models. For the Cardiac and Spleen datasets, negative samples are deleted from the small datasets with small objects to better segment the small objects. Moreover, considering the calculation effort, we set the resize of the original medical images and ground truths in the Spleen and Liver datasets to 480 × 480 [52]. Finally, for all datasets, there are 70% of the datasets for training, 10% for validation, and 20% for testing.

4.2. Baselines

In order to evaluate the performances of the proposed μ-Net, the four state-of-art deep-learning-based image segmentation methods FCN [23], U-Net [4], Attention U-Net [21], and ResUNet++ [30] are selected as baselines. The reasons for selecting these four methods as the baselines are as follows. (i) **FCN** is the first deep-learning-based end-to-end image segmentation model; (ii) **U-Net** is arguably the most widely adopted deep model for medical image segmentation, and it is also used as the backbone of the proposed μ-Net; (iii) **Attention U-Net** and **ResUNet++** are common variants of U-Net. Besides, we also select the four state-of-the-art deep supervision methods U-Net++ [17], U-Net3+ [13], M-Net [12], and MLDS-Net [14] as baselines. These deep supervision methods cover the state-of-the-art and most common deep supervision methods available: (i) **U-Net++**, **U-Net**3+, and **M-Net** are deep supervised models that up-samples the intermediate outputs to compare with the ground truths, and (ii) **MLDS-Net** is a deep supervised model that down-samples the ground truths to compare with the intermediate outputs.

Furthermore, to illustrate the effectiveness and necessity of the proposed Similarity Principle of Deep Supervision and tied-weight decoder proposed in μ-Net, ablation studies are further carried out, where several deeply supervised models using different pseudo-labels are introduced and evaluated. Specifically, the intermediate models are as follows: (i) **U-Net-Deep** is constructed by using semantically

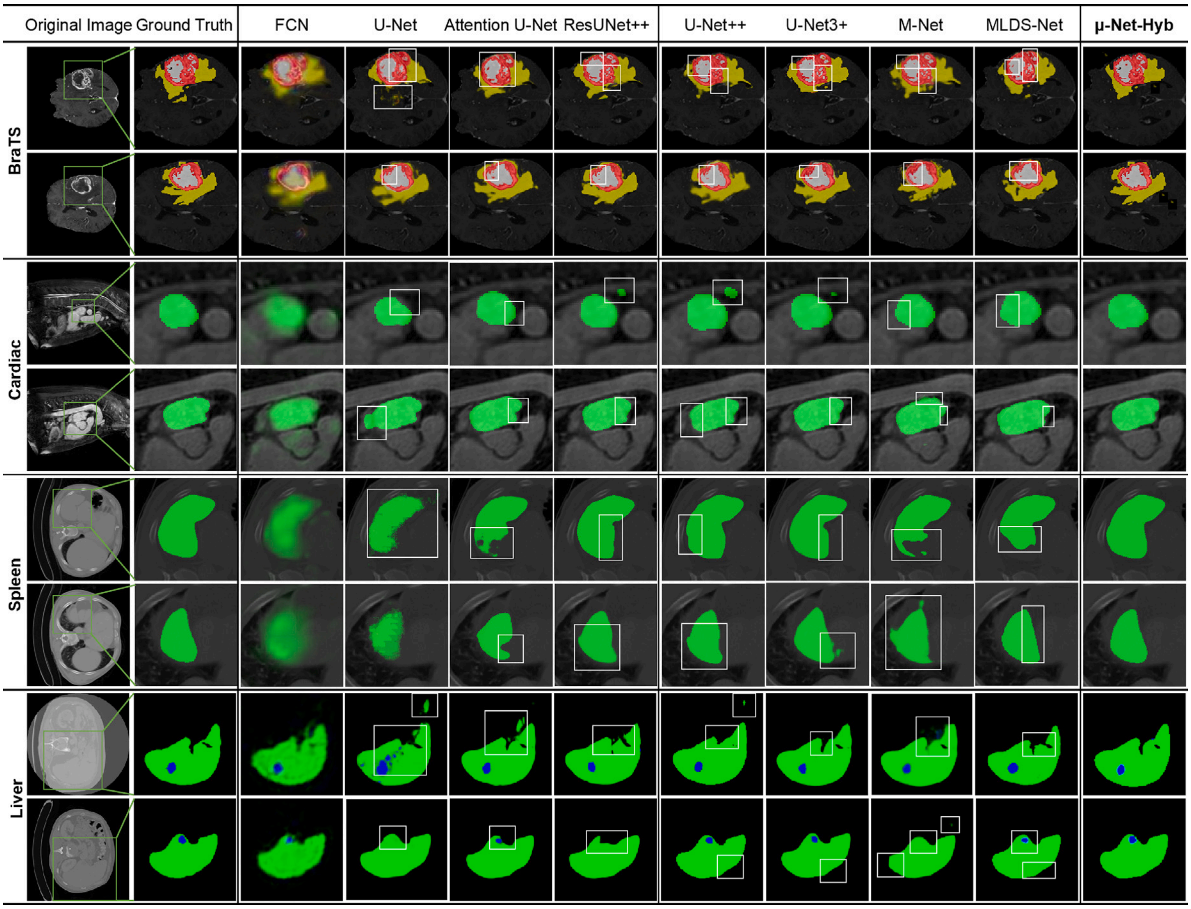


Fig. 2. Visualization of segmentation results of our proposed μ-Net-Hyb and the baselines on four datasets. The yellow, red, and white areas on BraTS are the whole tumor region, tumor core region, and tumor enhancement region, respectively; the green areas on Cardiac, Spleen, and Liver dataset are the segmented organs (i.e., heart, spleen, and liver); the blue areas on Liver is the resulting segmented areas of liver tumors; while the white boxes (i.e., rectangle boundary) are used to mark the wrongly segmented areas of the state-of-the-art baselines.

identical deep supervision onto U-Net; (ii) **U-Net-Auto** is based on U-Net using semantically similar deep supervision; and (iii) μ-Net is a model that incorporates a tied-weight decoder with U-Net, according to different training strategies, further divided into: (a) **μ-Net-Seg** is obtained by using identical semantic pseudo-labels into μ-Net; (b) **μ-Net-Rec** is μ-Net with reconstructive pseudo-labels; and (c) **μ-Net-Hyb** integrates the hybrid pseudo-labels with μ-Net. In addition to verifying the effectiveness of our deep supervision mechanisms, we also want to prove that our deep supervision mechanisms can be widely applicable to U-Net and its variants, so we further use U-Net++ as the backbone of the proposed strategies, resulting in **U-Net-Deep++**, **U-Net-Auto++**, **μ-Net++**, **μ-Net-Seg++**, **μ-Net-Rec++**, and **μ-Net-Hyb++**.

To demonstrate that the proposed Similarity Principle of Deep Supervision is a better choice for small object segmentation tasks than the state-of-the-art deep supervision, we further conduct some additional experiments to compare the Similarity Principle of Deep Supervision with two baselines: U-Net with Up and U-Net with Down. Specifically, we first refer to M-Net [12] to compare the intermediate outputs of the expansive path after up-sampling with the ground truths, and build a deep supervision model **U-Net with Up**; then, we refer to MLDS-Net [14] to down-sample the ground truths and compared with the intermediate outputs of the expansive path to form a deep supervision model **U-Net with Down**. Finally, this quantitative model is compared with U-Net-Deep and U-Net-Auto (U-Net with Similarity Principle of Deep Supervision), showing the superiority of our proposed Similarity Principle of Deep Supervision. Similarly, other additional experiments are also conducted to compare the tied-weight decoder with a variational autoencoder (VAE) decoder module. The module of VAE decoder, a state-of-the-art additional decoder module, is added to U-Net with Similarity Principle of Deep Supervision referring to [39], resulting **μ-Net using VAE decoder** is compared with μ-Net-Rec.

4.3. *Experimental settings*

Our experiments are implemented using the PyTorch framework[1] and run on NVIDIA TITAN XP 12 GB GPU (for larger image size in the Spleen dataset) and NVIDIA GeForce GTX 2080Ti GPU. The implementation details of the proposed μ-Net are shown as follows. The basic structure of μ-Net adopts 5-layer U-Net with k kernels of size 3×3, where $k = 64 \times 2^i$ (i indexes the down-sampling layer along with the decoder). Then, μ-Net and all the baselines are trained on the BraTS, Cardiac, and Spleen datasets using the Adam optimizer with mini-batch sizes of 4, 2, and 1, respectively. The learning rate is initialized as $\alpha_0 = 3e{-}4$ and progressively decreased for every three training epochs according to $\alpha = \alpha_0 * 0.9$. Moreover, we also use the *early-stop* mechanism on the validation set; specifically, the model stops training when the average of all metrics for training $2 * best_{epoch}$ (the epoch of average value of all metrics reaches the maximum value) no longer increases or reaches the maximum training epochs. Finally, we conduct grid search to investigate the effect of different loss weights on the μ-Net segmentation performance.

To show the effectiveness of our models, we use the Dice coefficient (DICE), Positive Predictive Value (PPV), Sensitivity (SEN), and

[1] https://pytorch.org/.

Table 2
Segmentation accuracies of μ-Net and the state-of-the-art baselines on four public databases. The best results are bold.

Anatomy	BraTS				Cardiac				Spleen				Liver			
	DICE	PPV	SEN	IoU	DICE	PPV	SEN	IoU	DICE	PPV	SEN	IoU	DICE	PPV	SEN	IoU
FCN [23]	0.6956	0.8382	0.7030	0.6457	0.8835	0.9028	0.8846	0.8532	0.8069	0.9058	0.8151	0.7588	0.8064	0.9330	0.8315	0.7871
U-Net [4]	0.7145	0.8408	0.7203	0.6651	0.9057	0.9045	0.9190	0.8754	0.8940	0.9133	0.8585	0.8362	0.8236	0.8988	0.8795	0.8022
Attention U-Net [21]	0.7148	0.8490	0.7411	0.6620	0.9183	0.9418	0.9162	0.8872	0.9310	0.9317	0.9522	0.8988	0.8580	0.9313	0.8858	0.8327
ResUNet++ [30]	0.7197	0.8629	0.7273	0.6701	0.9176	0.9249	0.9321	0.8893	0.9245	0.9296	0.9301	0.8862	0.8591	0.9191	0.8941	0.8339
U-Net++ [17]	0.7192	0.8426	0.7233	0.6657	0.9111	0.9089	0.9312	0.8763	0.9230	0.9187	0.8925	0.8722	0.8615	0.9236	0.8972	0.8362
U-Net3+ [13]	0.7292	0.8726	0.7396	0.6780	0.9304	0.9405	0.9364	0.9011	0.9440	0.9305	0.9369	0.8597	0.8700	0.9374	0.8932	0.8464
M-Net [12]	0.7163	0.8433	0.7229	0.6652	0.9135	0.9199	0.9209	0.8796	0.9243	0.9398	0.9328	0.8927	0.8310	0.9211	0.8654	0.8047
MLDS-Net [14]	0.7232	0.8618	0.7293	0.6715	0.9163	0.9227	0.9318	0.8870	0.9044	0.9116	0.8755	0.8473	0.8385	0.9236	0.8672	0.8130
μ-Net-Hyb (ours)	**0.7384**	**0.8950**	**0.7436**	**0.6880**	**0.9360**	**0.9489**	**0.9486**	**0.9061**	**0.9768**	**0.9767**	**0.9722**	**0.9617**	**0.8926**	**0.9395**	**0.9330**	**0.8681**
μ-Net-Hyb++ (ours)	**0.7472**	**0.8873**	**0.7646**	**0.6962**	**0.9413**	**0.9649**	**0.9490**	**0.9135**	**0.9786**	**0.9854**	**0.9748**	**0.9640**	**0.8953**	**0.9381**	**0.9496**	**0.8725**

Table 3
Training efficiencies of μ-Net and the state-of-the-art baselines on four public databases. The best results among all advanced U-Net are bold and the second best ones are underlined.

Anatomy	BraTS			Cardiac			Spleen			Liver		
	Avg.T(h)	Epoch	Tot.T(h)	Avg.T(h)	Epoch	Tot.T(h)	Avg.T(h)	Epoch	Tot.T(h)	Avg.T(h)	Epoch	Tot.T(h)
FCN [23]	0.1925	15	2.8877	0.0276	23	0.6360	0.0430	69	2.9701	0.2358	42	9.4330
U-Net [4]	0.1960	8	1.5679	0.0277	19	0.5269	0.0499	59	2.9462	0.1619	60	9.9216
Attention U-Net [21]	0.2667	39	10.4016	0.0466	51	2.3752	0.0612	65	3.9751	0.3392	49	12.9577
ResUNet++ [30]	0.6958	50	34.7905	0.0694	46	3.1913	0.1001	66	6.6076	0.5316	67	22.8983
U-Net++ [17]	0.5984	42	25.1322	0.0750	42	3.1486	0.1057	53	5.6005	0.3391	45	11.8250
U-Net3+ [13]	0.8438	32	27.0022	0.1088	30	3.2649	0.1522	52	7.9143	0.5480	35	19.1784
M-Net [12]	0.3073	30	9.2179	0.0405	50	2.0266	0.0848	83	7.0373	0.2538	51	12.9438
MLDS-Net [14]	0.2982	49	14.6107	0.0411	76	3.1248	0.0831	115	9.5576	0.2497	56	13.9832
μ-Net-Hyb (ours)	0.6470	**7**	**4.5289**	0.0841	**13**	**1.0934**	0.1459	**21**	**3.0639**	0.6145	**10**	**6.1450**
μ-Net-Hyb++ (ours)	1.0288	19	19.5476	0.1106	**10**	1.1056	0.1786	23	4.1078	0.8911	12	10.6932

Intersection over Union (IoU). Specifically, DICE (also known as F1) evaluates the overlap value between the outputs and the ground truths. It is the comprehensive evaluation metrics of PPV and SEN. PPV (also known as precision rate) measures the percentage of true positive samples of all predicted positive samples. SEN (also known as recall rate) has evaluated the probability that positive samples are correctly classified as positive. IoU is a standard of measuring the accuracy of corresponding objects in a specific dataset. Higher values of these metrics mean a better performance. Formally,

$$DICE = \frac{2 * TP + \epsilon}{T + P + \epsilon}, \qquad PPV = \frac{TP + \epsilon}{TP + FP + \epsilon},$$
$$SEN = \frac{TP + \epsilon}{TP + FN + \epsilon}, \qquad IoU = \frac{TP + \epsilon}{T + P - TP + \epsilon}, \tag{14}$$

where TP, FP, and FN are the number of true positive points, false positive points, and false negative points, respectively. T is the number of ground truth points of that class, and P is the number of predicted positive points. Finally, ϵ is a small constant to avoid zero division, which is set to 1 in our experiment. Besides, the total training time-cost (i.e., training efficiency) of a deep learning model ($Tot.T$) is determined by two factors: the time-cost of each training epoch $Avg.T$ (i.e., the time needed to train all samples once), and the number of training epochs needed for the deep model to converge $Epoch$; so we have $Tot.T = Avg.T \times Epoch$. In addition, we also use the p-value to measure the statistic significance of improvements.

4.4. Main results

The quantitative experimental results are shown in Tables 2 (accuracies) and 3 (efficiency), and examples of segmentation results of our proposed μ-Net and baselines on four datasets are shown in Fig. 2. To investigate the effectiveness of our proposed μ-Net, we conduct experiments on four datasets and compare the performance of μ-Net with four state-of-the-art segmentation baselines and four state-of-the-art deep supervision methods.

Generally, as shown in Table 2, our proposed μ-Net generally outperforms all the baselines, which proves that our proposed μ-Net achieves a more accurate medical image segmentation for small objects than the state-of-art image segmentation solutions. Specifically, we first find that deep-supervision models (i.e., U-Net++, U-Net3+, M-Net, and MLDS-Net) are generally better than FCN and U-Net on all datasets in terms of all metrics. This is not only because these models are structurally improved (i.e., U-Net++ and U-Net3+), but add deep supervision by additional supervision signals. This observation proves that deep models' segmentation performances can be improved by adding deep supervision using some additional supervision signals. Then, by comparing our μ-Net-Hyb and μ-Net-Hyb++ with FCN, U-Net, and its variants (i.e., Attention U-Net and ResUNet++), we observe that μ-Net-Hyb and μ-Net-Hyb++ consistently outperform FCN, U-Net, and its variants, which further proves that adding deep supervision can prevent the disappearance of the features of the important small objects and thus improve the segmentation accuracy of the model. Finally, Table 2 exhibits that the proposed μ-Net-Hyb and μ-Net-Hyb++ generally achieve a better segmentation performance than the existing deep-supervision-based models (i.e., U-Net++, U-Net3+, M-Net, and MLDS-Net) on all four datasets in terms of all metrics. This is because (i) our models utilize a semantic Similarity Principle of Deep Supervision on the basis of U-Net to advance and directly constrain the learning process of the intermediate layers; (ii) using semantically similar but not exactly identical deep supervision mechanism can achieve a better improvement in medical image segmentation than the state-of-the-art deep supervision mechanisms; and (iii) we also introduce a tied-weight decoder to strengthen the model's learning capability.

Furthermore, as shown in Table 3, to improve the segmentation performance, the existing deep supervision models introduce additional intermediate learning signals, which will cause the low-efficiency problem. On one hand, the model structures of deep supervised models M-Net and MLDS-Net are very similar to that of U-Net, and the only difference is adding some projection heads on the intermediate layers of U-Net's decoder to introduce some intermediate learning signals. By comparing the $Avg.T$ of M-Net and MLDS-Net with that of U-Net, we can find that $Avg.T$ of M-Net and MLDS-Net are always much higher than that of U-Net (e.g., 0.3073 and 0.2982 vs. 0.1925 on BraTS); the higher $Avg.T$ comes from both the additional projection operations needed to obtain the intermediate learning signals in feed-forward processes and the more complex objective functions in the M-Net and MLDS-Net models' backpropagation optimization processes. Therefore, these findings prove that introducing additional intermediate learning signals will greatly increase the $Avg.T$ in the existing deep supervised models. On the other hand, by comparing the $Epoch$ of M-Net and MLDS-Net with those of U-Net on BraTS and Cardiac (e.g., 30 and 49 vs. 8 on BraTS, 50 and 76 vs. 19 on Cardiac), we find that the problematic additional learning signals significantly increase the $Epoch$ of the existing deep supervised models in relatively easy tasks. As for the relatively difficult tasks, the results on Spleen and Liver also support our obverse: the $Epoch$ of M-Net and MLDS-Net is similar (higher on Spleen and slightly lower on Liver) to those of U-Net. Therefore, this phenomenon proves that additional intermediate learning signals in the existing deep supervised models will significantly increase the $Epoch$ in relatively easy segmentation tasks, while $Epoch$ in hard segmentation tasks is also similar to that of U-Net (the problematic additional learning signals have only limited promotion effect on the learning of the original features, so $Epoch$ is very unlikely to be reduced greatly). Consequently, all the above obverses fully demonstrate that the existing deep supervised models have a low learning efficiency problem: By introducing additional intermediate learning signals, the $Avg.T$ will be significantly increased and the $Epoch$ will also increase (in most cases) or remain similar (in even best case), so it is very likely that the $Tot.T$ is greatly increased. Please note that since U-Net++ and U-Net3+ have much more complex model structures (adding lots of additional convolutional operations on skip-connections) than U-Net (so as M-Net and MLDS-Net), their very high $Avg.T$ scores are not solely due to additional intermediate learning signals (but also due to much more complex structure); so, for fair comparison, we use M-Net and MLDS-Net instead of U-Net++ and U-Net3+ as illustration examples here. Then, we find that, despite of incorporating deep supervision and achieving superior accuracy performances, the total training time-cost of μ-Net-Hyb is much less than the state-of-the-art deep supervised segmentation models. This is because the tied-weight decoder in μ-Net dramatically speeds up the convergence process (the number of epochs needed for μ-Net-Hyb is only around one-third of those of the state-of-the-art deep supervision baselines), so even with a relatively high time-cost for each epoch, the total training time-cost of μ-Net-Hyb is much lower. Similar observations are also found for μ-Net-Hyb++; although its training time-cost is higher than μ-Net-Hyb due to the usage of a more complex backbone, the time is still much lower than those of the deep supervised baselines; this is also because of the very fast convergence process in the model training. Consequently, the findings in Table 3 sufficiently demonstrate that with the help of tied-weight decoders, μ-Net can overcome the low training efficiency problem and achieve not only accurate but also very efficient performances in medical image segmentation tasks.

Moreover, we also use p-value to measure the statistical significance of improvements of our methods w.r.t. the baselines. Specifically, we find that the p-values of our μ-Net-Hyb (resp., μ-Net-Hyb++) w.r.t. the baselines are between 0.2243 and 0.4236 (resp., between 0.1778 and 0.3684) on BraTS, between 0.0001 and 0.2803 (resp., between 0.00003 and 0.1402) on Cardiac, and between 0.0242 and 0.2008 (resp., between 0.0174 and 0.1532) on Liver, while the p-values of our μ-Net-Hyb (resp., μ-Net-Hyb++) w.r.t. the baselines are all equal or smaller than 0.0043 (resp., 0.0027) on Spleen. Consequently, we can find that the majority of the p-values are lower than 0.05; since, in the research areas of deep learning, it is impractical to always achieve statistically significant improvements, having p-values lower than 0.05 in most cases have been sufficient to prove that our proposed μ-Net can achieve significant improvements w.r.t the state-of-the-art segmentation baselines.

Finally, in order to visualize the superior performance of μ-Net-Hyb in medical image segmentation, Fig. 2 shows the segmentation results of four segmentation baselines (i.e., FCN, U-Net, Attention U-Net, and ResUNet++), four deep supervised models (i.e., U-Net++, U-Net3+, M-Net, and MLDS-Net), and μ-Net-Hyb on eight examples from four datasets. Specifically, the segmentation results of the brain tumor images at the first two rows of Fig. 2 show that: (i) the segmentation results of FCN and U-Net are very incorrect among the whole tumor region, tumor enhancement region, and tumor core region, and even the segmentation results are blurry; (ii) the segmentation results of the four deep supervised models are relatively better than FCN and U-Net, but their segmentation results for the tumor core region are still not ideal (such as white boxes); and (iii) the segmentation performance of μ-Net-Hyb is much better than the four segmentation baselines and the four deep supervised models, its segmentation results for the brain tumors are all very close to the ground truths. Similarly, from the segmentation results of the cardiac images at the third and fourth rows of Fig. 2, we have the following observations: (i) FCN and U-Net cannot correctly recognize and segment the heart; (ii) the four deep supervised models are better than FCN and U-Net, but their performances in segmenting the edge areas of the heart are not satisfactory and even over-segmentation; and (iii) the segmentation results of the proposed μ-Net-Hyb are best among all four segmentation baselines and four deep supervised models and are all closest to the ground truths. Similar observations are also found for the spleen images, where μ-Net-Hyb is the only model that correctly segments the spleen with smooth edges (such as the white boxes). Similar observations are also found for the liver images, where μ-Net-Hyb is the only model that correctly segments the liver and liver tumor (such as the white boxes). Therefore, these visualized observations greatly demonstrate again that by the proposed semantically similar but not completely consistent deep supervision mechanism, μ-Net-Hyb, remedies the drawbacks of the existing deep segmentation models and deep supervision models, and achieves a much better performance in medical image segmentation tasks, especially for small objects and objects with complex boundary details.

4.5. Ablation studies

To further investigate the effectiveness and necessity of the Similarity Principle of Deep Supervision and tied-weight decoder, ablation studies are conducted with five deep supervision models that use different pseudo-labels based on U-Net (resp., U-Net++), i.e., U-Net-Deep (resp., U-Net-Deep++), U-Net-Auto (resp., U-Net-Auto++), μ-Net-Seg (resp., μ-Net-Seg++), μ-Net-Rec (resp., μ-Net-Rec++), and μ-Net-Hyb (resp., μ-Net-Hyb++). The corresponding experimental results are shown in Tables 4 (accuracy) and 5 (efficiency).

In Table 4, all five deep supervision models outperform U-Net (resp., U-Net++) over all four datasets in terms of all metrics, which proves that the proposed deep supervision mechanisms are all effective to improve the performance of U-Net (resp., U-Net++) in medical image segmentation tasks. Specifically, we first compare the results of U-Net (resp., U-Net++) with the five deep supervision models based on U-Net (resp., U-Net++), where the deep supervision models outperform U-Net (resp., U-Net++) on all datasets in terms of all metrics. This is because deep supervision can early and directly enhance the feature learning capability of the deep layers by adding additional supervision signals in the expansive paths. Then, it is observed that U-Net-Auto (resp., U-Net-Auto++) and μ-Net-Rec (resp., μ-Net-Rec++) consistently outperform U-Net-Deep (resp., U-Net-Deep++) and μ-Net-Seg (resp., μ-Net-Seg++) in Table 4. This is because the former uses pseudo-labels that are semantically similar to the intermediate outputs, rather than the completely identical pseudo-labels used by the latter. Therefore, this proves the correctness of the Similarity Principle of Deep Supervision, which can make the generated pseudo-labels retain the diverse image details as much as possible, thereby guiding intermediate predictions to retain more details. Furthermore, we note that μ-Net-Seg (resp., μ-Net-Seg++) and μ-Net-Rec (resp., μ-Net-Rec++) are always better than U-Net-Deep

Table 4
Ablation studies in segmentation accuracy. imp. presents the performance improvements of μ-Net-Hyb (resp., μ-Net-Hyb++) with respect to U-Net (resp., U-Net++). The best results are bold.

Anatomy	BraTS				Cardiac				Spleen				Liver			
	DICE	PPV	SEN	IoU	DICE	PPV	SEN	IoU	DICE	PPV	SEN	IoU	DICE	PPV	SEN	IoU
U-Net [4]	0.7145	0.8408	0.7203	0.6651	0.9057	0.9045	0.9190	0.8754	0.8940	0.9133	0.8585	0.8362	0.8530	0.9268	0.8832	0.8289
U-Net-Deep	0.7245	0.8701	0.7315	0.6737	0.9252	0.9386	0.9327	0.8951	0.9246	0.9301	0.9311	0.8524	0.8539	0.9329	0.8838	0.8294
U-Net-Auto	0.7268	0.8713	0.7348	0.6764	0.9307	0.9401	0.9361	0.9013	0.9291	0.9315	0.9426	0.8664	0.8543	0.9330	0.8845	0.8351
μ-Net-Seg	0.7319	0.8774	0.7400	0.6810	0.9308	0.9435	0.9365	0.9015	0.9681	0.9401	0.9523	0.9065	0.8546	0.9333	0.8922	0.8371
μ-Net-Rec	0.7340	0.8807	0.7434	0.6836	0.9332	0.9455	0.9393	0.9024	0.9744	0.9652	0.9605	0.9525	0.8623	0.9347	0.8977	0.8375
μ-Net-Hyb	**0.7384**	**0.8950**	**0.7436**	**0.6880**	**0.9360**	**0.9489**	**0.9486**	**0.9061**	**0.9768**	**0.9767**	**0.9722**	**0.9617**	**0.8926**	**0.9395**	**0.9330**	**0.8681**
imp.	0.0239	0.0542	0.0233	0.0229	0.0303	0.0444	0.0296	0.0307	0.0828	0.0634	0.1137	0.1255	0.0396	0.0127	0.0498	0.0392
U-Net++ [17]	0.7192	0.8426	0.7233	0.6657	0.9111	0.9089	0.9312	0.8763	0.9230	0.9187	0.8925	0.8722	0.8615	0.9236	0.8972	0.8362
U-Net-Deep++	0.7341	0.8705	0.7540	0.6820	0.9315	0.9395	0.9331	0.8994	0.9284	0.9651	0.9394	0.8827	0.8696	0.9253	0.8977	0.8450
U-Net-Auto++	0.7368	0.8721	0.7609	0.6851	0.9319	0.9411	0.9381	0.8998	0.9345	0.9665	0.9478	0.8921	0.8743	0.9343	0.9027	0.8465
μ-Net-Seg++	0.7391	0.8795	0.7617	0.6883	0.9336	0.9458	0.9392	0.9023	0.9720	0.9838	0.9692	0.9561	0.8757	0.9397	0.9042	0.8466
μ-Net-Rec++	0.7420	0.8820	0.7637	0.6903	0.9368	0.9474	0.9465	0.9041	0.9782	0.9847	0.9747	0.9634	0.8783	0.9404	0.9070	0.8546
μ-Net-Hyb++	**0.7472**	**0.8873**	**0.7646**	**0.6962**	**0.9413**	**0.9649**	**0.9490**	**0.9135**	**0.9786**	**0.9854**	**0.9748**	**0.9640**	**0.8953**	**0.9441**	**0.9214**	**0.8725**
imp.	0.0280	0.0447	0.0413	0.0305	0.0302	0.0560	0.0178	0.0372	0.0556	0.0667	0.0823	0.0918	0.0338	0.0205	0.0242	0.0363

Table 5
Ablation studies in training efficiencies. The best results are bold and the second best ones are underlined.

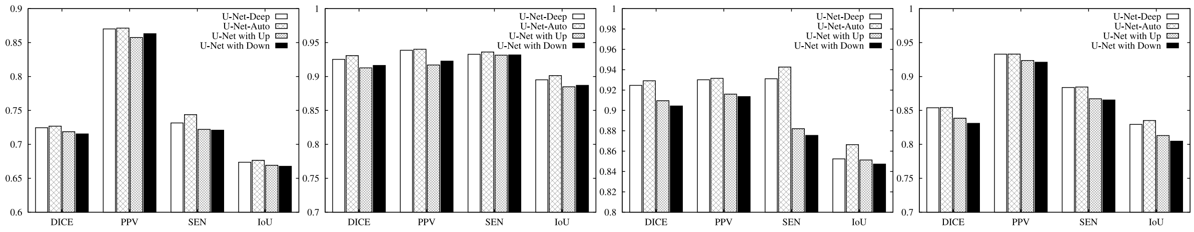
Anatomy	BraTS			Cardiac			Spleen			Liver		
	Avg.T(h)	Epoch	Tot.T(h)	Avg.T(h)	Epoch	Tot.T(h)	Avg.T(h)	Epoch	Tot.T(h)	Avg.T(h)	Epoch	Tot.T(h)
U-Net [4]	0.1960	8	**1.5679**	0.0277	19	**0.5269**	0.0499	59	**2.9462**	0.1619	60	9.9216
U-Net-Deep	0.3778	14	5.2888	0.0555	39	2.1649	0.1076	55	5.9202	0.6589	31	21.9312
U-Net-Auto	0.6386	12	7.6635	0.0795	23	1.8286	0.1400	45	6.2980	0.8831	15	13.0363
μ-Net-Seg	0.4769	11	5.2463	0.0748	23	1.7211	0.1295	39	5.0524	0.4962	17	8.4356
μ-Net-Rec	0.5696	8	4.5569	0.0761	19	1.4450	0.1253	39	4.8880	0.5222	14	7.3112
μ-Net-Hyb	0.6470	**7**	4.5289	0.0841	**13**	1.0934	0.1459	**21**	3.0639	0.6145	**10**	**6.1450**
U-Net++ [17]	0.5984	42	25.1322	0.0750	42	3.1486	0.1057	53	5.6005	0.3391	45	11.8250
U-Net-Deep++	0.8009	69	55.2636	0.0804	35	2.8157	0.1403	38	5.3312	1.4708	18	26.2477
U-Net-Auto++	0.9990	42	41.9564	0.0945	32	3.0237	0.1769	32	5.6599	1.4315	13	18.3910
μ-Net-Seg++	0.9067	41	37.1746	0.0853	31	2.6452	0.1609	32	5.1482	0.8051	25	20.1275
μ-Net-Rec++	0.9472	40	37.8878	0.0919	21	1.9304	0.1647	25	4.1166	0.8247	27	22.2669
μ-Net-Hyb++	1.0288	**19**	**19.5476**	0.1106	**10**	**1.1056**	0.1786	**23**	**4.1078**	0.8911	**12**	**10.6932**

(resp., U-Net-Deep++) and U-Net-Auto (resp., U-Net-Auto++), because the former introduces a tied-weight decoder to add additional reconstructed supervision signals. Finally, we find that μ-Net-Hyb (resp., μ-Net-Hyb++) consistently outperforms the other four deep supervision models, because it combines μ-Net-Seg (resp., μ-Net-Seg++) and μ-Net-Rec (resp., μ-Net-Rec++) with different detailed information that is beneficial to the deep segmentation models, which can further expand the diversity of detailed information of pseudo-labels, thereby better guiding intermediate predictions to further improve the segmentation performance.

Then, in Table 5, we observe that U-Net-Deep and U-Net-Auto can improve the model's convergence performance. This can be demonstrated by comparing the $Epoch$ of U-Net-Deep in Table 5 and those of M-Net and MLDS-Net in Table 3 (similar to U-Net-Auto). Specifically, the model structure of the proposed strategy U-Net-deep is very similar to those of M-Net and MLDS-Net (i.e., U-Net plus side projections or outputs in the decoders), the only difference between U-Net-Deep and M-Net and MLDS-Net is that: in M-Net and MLDS-Net, there exists semantic difference problem between their intermediate masks and intermediate predictions used to construct their intermediate learning signals, while the intermediate learning signals in U-Net-deep does not have the semantic difference problem. The results in Tables 3 and 5 show that by resolving the semantic difference problem, the $Epoch$ of U-Net-Deep is much lower than those of M-Net and MLDS-Net: 14 vs. 30 and 49 on BraTS, 39 vs. 50 and 57 on Cardiac, 55 vs. 83 and 115 on Spleen, and 31 vs. 51 and 56 on the Liver. Consequently, in Tables 3 and 5, the $Tot.T$ of U-Net-Deep is much lower than those of M-Net and MLDS-Net on BraTS, Cardiac, and Spleen datasets. However, we also need to notice that solely relying on this improvement is not enough to fully resolve the low-efficiency problem: when the increase of $Avg.T$ overwhelms the decrease of $Epoch$, the $Tot.T$ of U-Net-Deep may also higher than those of M-Net and MLDS-Net (just like the case on Liver dataset, where $Tot.T$ is 21.9312 for U-Net-Deep, while 12.9438 and 13.9832 for M-Net and MLDS-Net, respectively).

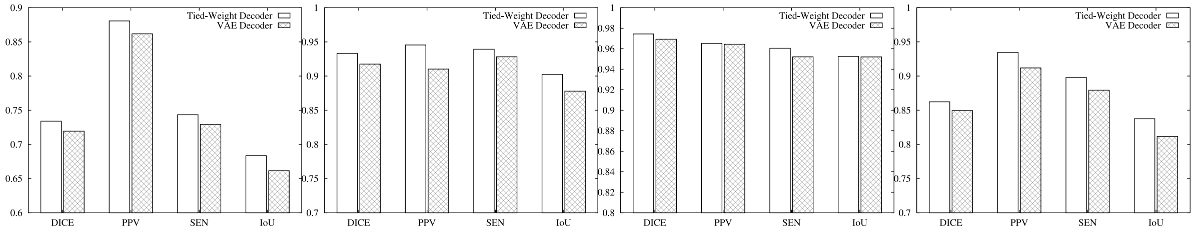
Therefore, we propose μ-Net to use a tied-weight decoder to dramatically further decrease the numbers of convergence epochs and ensure that the decrease of epochs is always much higher than the increase of time-cost per epochs. Specifically, by using a tied-weight decoder, the numbers of convergence epochs of three versions of μ-Net (i.e., μ-Net-Seg, μ-Net-Rec, and μ-Net-Hyb) are all much lower than those of U-Net-Deep and U-Net-Auto on all datasets in Table 5. In addition, since the decrease of epochs is always much higher than the increase of time-cost per epoch, the total training time of μ-Net-Hyb constantly outperforms those of all state-of-the-art advanced U-Net models (i.e., the conventional deep supervised models and Attention U-Net and ResUNet++) on all datasets. Consequently, all the above obverses fully demonstrate that the explored U-Net-deep, U-Net-Auto and μ-Net can solve the low learning efficiency problem. We have to emphasize again that U-Net-deep and U-Net-Auto address the low efficiency in most (but not all) situations, and only μ-Net fully addresses the low-efficiency problem.

Furthermore, by comparing the results of U-Net-Auto (resp., μ-Net-Rec) with those of U-Net-Deep (resp., μ-Net-Seg) in Tables 4 and 5 of the revised manuscript, we can find that U-Net-Auto (resp., μ-Net-Rec) constantly outperforms U-Net-Deep (resp., μ-Net-Seg) in both segmentation accuracy and training efficiency. We believe this is because the similar but not identical convolution and deconvolution procedures make the generated intermediate masks and intermediate predictions in U-Net-Auto (resp., μ-Net-Rec) not only contain semantics at the same depth but also capable of learning more diverse and task-specific features than those generated by identical procedures in U-Net-Deep (resp., μ-Net-Seg). According to these observations, we thus summarize



(a) BraTS (b) Cardiac (c) Spleen (d) Liver

Fig. 3. Similarity Principle of Deep Supervision vs. the state-of-the-art deep supervisions.



(a) BraTS (b) Cardiac (c) Spleen (d) Liver

Fig. 4. μ-Net using tied-weight decoder vs. μ-Net using VAE decoder.

the Similarity Principle of Deep Supervision, which can be used to measure the quality of the intermediate learning signals used in deep supervised models, and will be beneficial for future research in deep supervised segmentations.

Finally, to show that when using U-Net++ as the backbone, the proposed strategies will also achieve similar improvements to those using U-Net as the backbone, and in turn, prove the high adaptability and scalability of our proposed strategies, we apply the proposed strategies on U-Net++ and show the corresponding ablation study results in Tables 4 and 5. By measuring the improvements (denoted imp. in tables) of μ-Net-Hyb and μ-Net-Hyb++ w.r.t. U-Net and U-Net++, respectively, we can discover that applying μ-Net on U-Net and U-Net++ actually obtain the similar extent of improvements in segmentation accuracies, while the total time-costs of μ-Net-Hyb and μ-Net-Hyb++ are also similar (i.e., slightly higher or lower) to those of U-Net and U-Net++, respectively. Therefore, the good adaptability and scalability of our proposed strategies are proven. In addition, it is also observed that μ-Net-Hyb++ has longer training time-costs than μ-Net-Hyb, this is because its backbone U-Net++ has much higher time-costs than μ-Net-Hyb's backbone U-Net (i.e., 25.1322 vs. 1.5679 on BraTS, 3.1486 vs. 0.5269 on Cardiac, 5.6005 vs. 2.9462 on Spleen, and 11.8250 vs. 9.9216 on Liver); therefore, although μ-Net-Hyb++ has applied tied-weight decoder to significantly decrease the training time-cost, it is unrealistic to fully cover the efficiency gap caused by the different backbones.

4.6. Similarity Principle of Deep Supervision vs. the state-of-the-art deep supervision mechanisms

Further experiments are conducted to compare our proposed Similarity Principle of Deep Supervision with the current state-of-the-art deep supervision mechanisms, namely, U-Net-Deep, U-Net-Auto, U-Net with Up (refer to M-Net [12] to up-sample the intermediate predictions and compared with the ground truths), and U-Net with Down (refer to MLDS-Net [14] to down-sample the ground truths and compared with the ground truths), where the different deep supervision mechanisms are respectively incorporated with U-Net to show their different capabilities in enhancing U-Net's performances in medical image segmentation with small objects and objects with complex boundary details. The corresponding experimental results are depicted in Fig. 3.

Generally, as shown in Fig. 3, the model of incorporating U-Net with our Similarity Principle of Deep Supervision can achieve much better performance improvements than using the state-of-the-art models of U-Net with Up, U-Net with Down, and semantically identical deep supervision, in terms of all metrics on all four datasets. This finding thus proves that the Similarity Principle of Deep Supervision is a better choice for small objects and objects with complex boundary details segmentation tasks than the state-of-the-art deep supervision mechanisms. Specifically, both U-Net-Deep and U-Net-Deep outperform the models of U-Net with Up and U-Net with Down. This shows that compared with existing deep supervision models, the semantic difference between intermediate predictions and corresponding labels in our deep supervision models is smaller, which can improve the model's learning ability and speed up the model's convergence performance. Furthermore, we note that U-Net-Auto is much better than U-Net-Deep in all the cases, which is because U-Net-Auto uses pseudo-labels with high semantic similarity rather than semantic identity for supervision, preserving more and richer semantic information. In summary, these findings clearly demonstrate the effectiveness and reasonableness of the proposed Similarity Principle of Deep Supervision in achieving better medical image segmentation performances than the state-of-the-art deep supervision mechanisms.

4.7. Tied-weight decoder vs. the state-of-the-art VAE decoder

Similarly, to investigate the influence of different additional decoders on medical images for small objects and objects with complex boundary information segmentation, we further conduct experiments to compare our introduced tied-weight decoder with the state-of-the-art and common VAE decoder module with U-Net, namely, compare μ-Net using a tied-weight decoder (i.e., μ-Net-Rec) with μ-Net using VAE decoder (i.e., adding VAE decoder to U-Net after integrating the Similarity Principle of Deep Supervision), the results are depicted in Fig. 4. Generally, the results show that the model of combining U-Net and our proposed Similarity Principle of Deep Supervision post-fusion tied-weight decoder (denoted μ-Net with tied-weight decoder) generally outperforms the model of combining U-Net and the Similarity Principle of Deep Supervision with VAE decoder module in terms of all metrics on all four datasets, which prove that the tied-weight decoder can achieve a more accurate medical image segmentation with small objects and objects with complex boundary details than the state-of-the-art VAE decoder module.

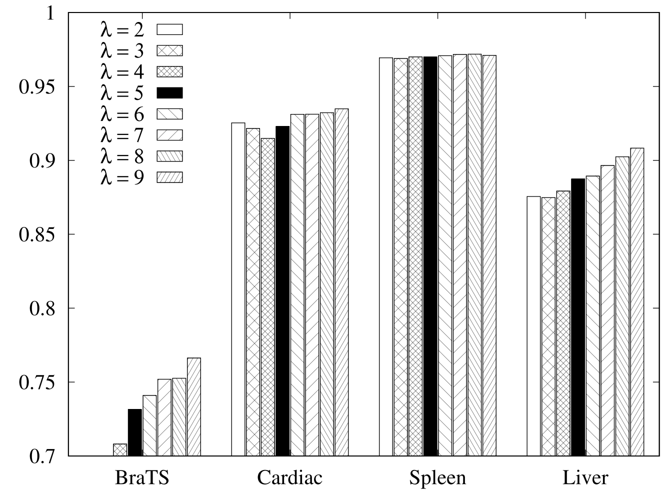


Fig. 5. Performance of μ-Net under different settings of loss-weights.

4.8. *Effect of varying loss-weights*

As shown in Eq. (13), the final loss function of our optimal segmentation model μ-Net-Hyb is affected by the weights of the individual loss functions. Therefore, the values of α, β, a, b, γ, ξ, λ, ω_i, η, and σ_i all affect the training quality of the model and the final segmentation performance. We consider that each deep layer in the expansive path to be equally important and therefore set both ω_i and σ_i to 1. Similarly, we also believe that whether it is the pseudo-labels generated by the U-Net decoder for supervision or the pseudo-labels generated by the tied-weight decoder for supervision, both contain useful information for segmentation, and assume that they are of equal weight, so set $\gamma = \xi = 1$. Consequently, experiments are conducted to investigate the effect of different loss weights on the model's training quality in terms of the average value of the four evaluation metrics, i.e., $avg\_m = \frac{DICE+PPV+SEN+IoU}{4}$.

We select all the values of loss weights incrementally from 2 to 9 with a step of 1, and the extreme weights 1 and 10 are not shown here. Generally, in Fig. 5, we observe that μ-Net-Hyb obtains the largest segmented evaluation metrics when the loss weights α, β, a, b, λ, and η are set 5, 1, 1, 1, 9, and 6 on the BraTS dataset; the loss weights α, β, a, b, λ, and η are set 7, 1, 4, 1, 9, and 6 on the Cardiac dataset; the loss weights α, β, a, b, λ, and η are set 3, 5, 2, 2, 8, and 6 on the Spleen dataset; the loss weights α, β, a, b, λ, and η are set 6, 1, 3, 1, 9, and 6 on the Spleen dataset, respectively, which are thus used as the final selected values. Moreover, the results in Fig. 5 show as another important finding that the segmentation metrics also improve with the increase of the value of λ within a certain range. This finding thus further proves that rational using of the Similarity Principle of Deep Supervision instead of directly using the highest weighted Similarity Principle of Deep Supervision can improve the segmentation performance of μ-Net-Hyb.

4.9. *Strategies selection in practical usage*

As demonstrated in our experimental studies, although all proposed strategies (U-Net-Deep, U-Net-Auto, and μ-Net) are applicable and achieve satisfactory performances on all kinds of medical image segmentation tasks, our experimental results show that μ-Net generally outperforms U-Net-Deep and U-Net-Auto in terms of both segmentation accuracy and training efficiency among all kinds of medical image segmentation tasks. Therefore, we suggest giving priority to using μ-Net on the clinical usages, no matter what kinds of medical images.

In addition, as stated in Section 1, μ-Net not only can use U-Net as the backbone but can also be used in most of the advanced U-Net models to further improve their accuracy effectively; and our experimental studies also demonstrate that by using U-Net++ as the backbone segmentation model, the resulting μ-Net-Hyb++ can further achieve better segmentation accuracies than U-Net based μ-Net-Hyb with the cost of, however, higher training time-cost (due to the much more complicated structure in the backbone); therefore, if the specific piratical segmentation task has high requirements for segmentation accuracy and has sufficient computing resources, the users can utilize μ-Net-Hyb++ to achieve more precious segmentations. Similarly, if one wants to further improve the segmentation accuracies, more complicated and advanced U-Net-based segmentation models, e.g., U-Net3+ [13] and ResUNet++ [30], can be further used as the backbone of μ-Net to achieve this aim.

5. Conclusion and future works

In this work, we identify the problem of existing deep supervision mechanisms, namely, semantic difference problem and low learning efficiency problem, and propose some deep supervised models to remedy the problems and achieve a more effective and efficient medical image segmentation. Specifically, we first propose U-Net-Deep and U-Net-Auto to overcome the semantic difference problem. Then, we further propose μ-Net, which designs a Similarity Principle of Deep Supervision to improve the model's segmentation performance and introduces a tied-weight decoder to accelerate the model's convergence performance. Finally, we explore three different types of μ-Net-based deep supervision strategies. Extensive experimental studies are conducted on three real-world medical image segmentation datasets with U-Net and U-Net++ as backbones, and the results show that the proposed μ-Net can significantly outperform the state-of-the-art image segmentation solutions in medical image segmentation tasks in terms of all metrics, and the Similarity Principle of Deep Supervision and tied-weight decoder are all effective and essential for μ-Net to achieve superior segmentation performance and are applicable to all U-Net variants.

Despite achieving generally superior performance in medical image segmentation tasks, we also observe in the experimental results that the performance of all segmentation models, including μ-Net, on the BraTS dataset are much worse than those on the Cardiac, Spleen, and Liver datasets. This is because the shape and appearance of brain tumors in medical images are much more varied than those of the heart and spleen, and the boundary information is more complex, so it is more difficult for the deep model to learn its morphological features. Therefore, it is interesting future work to further improve the deep supervision mechanisms in μ-Net to solve this problem, such as adding some supervision signals appropriately in the extracting path, so that μ-Net is more applicable in the segmentation tasks of segmenting objects with complex boundary information.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under the grants 62276089 and 61906063, by the Natural Science Foundation of Hebei Province, China, under the grant F2021202064, by the Key Research and Development Project of Hainan Province, China, under the grant ZDYF2022SHFZ015, by the Natural Science Foundation of Hainan Province, China, under the grant 821RC1131, and by the China Scholarship Council Fund, China, under the number 202206700006. This work was also supported by the Hainan Province Clinical Medical Center, China, under the grant QWYH202175, and by the AXA Research Fund.

References

[1] E. Gibson, W. Li, C. Sudre, L. Fidon, D.I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, et al., NiftyNet: A deep-learning platform for medical imaging, Comput. Methods Programs Biomed. 158 (2018) 113–122.

[2] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, Q. Sun, Deep learning for image-based cancer detection and diagnosis: A survey, Pattern Recognit. 83 (2018) 134–149.

[3] A.V. Dalca, J. Guttag, M.R. Sabuncu, Anatomical priors in convolutional networks for unsupervised biomedical segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9290–9299.

[4] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.

[5] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, C. Pal, The importance of skip connections in biomedical image segmentation, in: Proceedings of the International Workshop on Deep Learning in Medical Image Analysis, International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, 2016, pp. 179–187.

[6] D. Lachinov, E. Vasiliev, V. Turlapov, Glioma segmentation with cascaded U-Net, in: Proceedings of the MICCAI Workshop: Brainlesion on Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 2018, pp. 189–198.

[7] S. Qamar, H. Jin, R. Zheng, P. Ahmad, M. Usama, A variant form of 3D-UNet for infant brain segmentation, Future Gener. Comput. Syst. 108 (2020) 613–623.

[8] X. Wang, S. Yang, Y. Fang, Y. Wei, M. Wang, J. Zhang, X. Han, SK-UNet: An improved U-Net model with selective kernel for the segmentation of LGE cardiac MR images, IEEE Sens. J. 21 (10) (2021) 11643–11653.

[9] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.-A. Heng, H-DenseUNet: Hybrid densely connected U-Net for liver and tumor segmentation from CT volumes, IEEE Trans. Med. Imaging 37 (12) (2018) 2663–2674.

[10] Z. Honghan, D.C. Liu, L. Jingyan, P. Liu, H. Yin, Y. Peng, RMS-SE-UNet: A segmentation method for tumors in breast ultrasound images, in: Proceedings of the International Conference on Computer and Communication Systems, 2021, pp. 328–334.

[11] X. Xiao, S. Lian, Z. Luo, S. Li, Weighted Res-UNet for high-quality retina vessel segmentation, in: Proceedings of the International Conference on Information Technology in Medicine and Education, 2018, pp. 327–331.

[12] H. Fu, J. Cheng, Y. Xu, D.W.K. Wong, J. Liu, X. Cao, Joint optic disc and cup segmentation based on multi-label deep network and polar transformation, IEEE Trans. Med. Imaging 37 (7) (2018) 1597–1605.

[13] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, UNet 3+: A full-scale connected U-Net for medical image segmentation, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 1055–1059.

[14] S. Reiß, C. Seibold, A. Freytag, E. Rodner, R. Stiefelhagen, Every annotation counts: Multi-label deep supervision for medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9532–9542.

[15] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2015, pp. 562–570.

[16] N. Ibtehaz, M.S. Rahman, MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87.

[17] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A nested U-Net architecture for medical image segmentation, in: Proceedings of the MICCAI Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018, pp. 3–11.

[18] Z. Xu, S. Liu, D. Yuan, L. Wang, J. Chen, T. Lukasiewicz, Z. Fu, R. Zhang, ω-Net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution, Neurocomputing (2022).

[19] L. Song, K. Geoffrey, H. Kaijian, Bottleneck feature supervised U-Net for pixel-wise liver and tumor segmentation, Expert Syst. Appl. 145 (2020) 113–131.

[20] Z. Xu, T. Lukasiewicz, C. Chen, Y. Miao, X. Meng, Tag-aware personalized recommendation using a hybrid deep model, in: Proceedings of the International Joint Conferences on Artificial Intelligence, 2017, pp. 3196–3202.

[21] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention U-Net: Learning where to look for the pancreas, 2018, ArXiv Preprint, ArXiv:1804.03999.

[22] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, Z. Xu, Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation, Med. Image Anal. 83 (2023) 102656.

[23] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[24] A. Ben-Cohen, I. Diamant, E. Klang, M. Amitai, H. Greenspan, Fully convolutional network for liver segmentation and lesions detection, in: Proceedings of the International Conference on MICCAI Deep Learning and Data Labeling for Medical Applications Workshop, 2016, pp. 77–85.

[25] X. Zhou, R. Takayama, S. Wang, T. Hara, H. Fujita, Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method, Med. Phys. 44 (10) (2017) 5221–5233.

[26] C. Wang, Z. Zhao, Q. Ren, Y. Xu, Y. Yu, Dense U-Net based on patch-based learning for retinal vessel segmentation, Entropy 21 (2) (2019).

[27] Z. Xu, C. Qi, G. Xu, Semi-supervised attention-guided cyclegan for data augmentation on medical images, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, 2019, pp. 563–568.

[28] D. Yuan, Y. Liu, Z. Xu, Y. Zhan, J. Chen, T. Lukasiewicz, Painless and accurate medical image analysis using deep reinforcement learning with task-oriented homogenized automatic pre-processing, Comput. Biol. Med. 153 (2023) 106487.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the International Conference on Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[30] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H.D. Johansen, ResUNet++: An advanced architecture for medical image segmentation, in: Proceedings of the IEEE International Symposium on Multimedia, 2019, pp. 225–2255.

[31] Y. Xu, Y. Li, M. Liu, Y. Wang, M. Lai, I. Eric, C. Chang, Gland instance segmentation by deep multichannel side supervision, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 496–504.

[32] K. Chen, M. Weinmann, X. Sun, M. Yan, S. Hinz, B. Jutzi, M. Weinmann, Semantic segmentation of aerial imagery via multi-scale shuffling convolutional neural networks with deep supervision, ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci. 4 (1) (2018).

[33] L. Wang, B. Wang, Z. Xu, Tumor segmentation based on deeply supervised multi-scale U-Net, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, 2019, pp. 746–749.

[34] Y. Liu, M.-M. Cheng, X.-Y. Zhang, G.-Y. Nie, M. Wang, DNA: Deeply supervised nonlinear aggregation for salient object detection, IEEE Trans. Cybern. (2021).

[35] Y. Zhang, A.C. Chung, Deep supervision with additional labels for retinal vessel segmentation task, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 83–91.

[36] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, P.-A. Heng, 3D deeply supervised network for automatic liver segmentation from CT volumes, in: Proceeding of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 149–157.

[37] B. Li, C. Wu, J. Chi, X. Yu, G. Wang, A deeply supervised convolutional neural network for brain tumor segmentation, in: Proceedings of the Chinese Control Conference, 2020, pp. 6262–6267.

[38] H. Liu, Q. Li, I. Wang, A deep-learning model with learnable group convolution and deep supervision for brain tumor segmentation, Math. Probl. Eng. 2021 (2021).

[39] A. Myronenko, 3D MRI brain tumor segmentation using autoencoder regularization, in: Proceedings of the MICCAI Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 2018, pp. 311–320.

[40] S.A. Kohl, B. Romera-Paredes, K.H. Maier-Hein, D.J. Rezende, S. Eslami, P. Kohli, A. Zisserman, O. Ronneberger, A hierarchical probabilistic U-Net for modeling multi-scale ambiguities, 2019, ArXiv Preprint, ArXiv:1905.13077.

[41] R.J. Araújo, J.S. Cardoso, H.P. Oliveira, A deep learning design for improving topology coherence in blood vessel segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 93–101.

[42] Z. Xu, T. Li, Y. Liu, Y. Zhan, J. Chen, T. Lukasiewicz, PAC-Net: Multi-pathway FPN with position attention guided connections and vertex distance IoU for 3D medical image detection, Front. Bioeng. Biotechnol. 11 (2023) 1049555.

[43] J. Fu, Y. Wu, T. Mei, J. Wang, H. Lu, Y. Rui, Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1985–1993.

[44] Q. Yu, Y. Xia, L. Xie, E.K. Fishman, A.L. Yuille, Thickened 2D networks for efficient 3D medical image segmentation, 2019, ArXiv Preprint, ArXiv:1904.01150.

[45] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, J. Fürnkranz, Large-scale multi-label text classification—revisiting neural networks, in: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2014, pp. 437–452.

[46] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: Proceedings of the International Conference on 3D Vision, 2016, pp. 565–571.

[47] A. Obot, O. Simeon, J. Afolayan, Comparative analysis of path loss prediction models for urban macrocellular environments, Niger. J. Technol. 30 (3) (2011) 50–59.

[48] L. Deng, Three classes of deep learning architectures and their applications: A tutorial survey, APSIPA Trans. Signal Inf. Process. 57 (2012) 58.

[49] J. Wang, T. Lukasiewicz, X. Hu, J. Cai, Z. Xu, RSG: A simple but effective module for learning imbalanced datasets, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3784–3793.

[50] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024.

[51] A.L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B.A. Landman, G. Litjens, B. Menze, et al., A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019, ArXiv Preprint, ArXiv:1902.09063.

[52] S. Singh, K. Ho-Shon, S. Karimi, L. Hamey, Modality classification and concept detection in medical images using deep transfer learning, in: Proceedings of the International Conference on Image and Vision Computing New Zealand, 2018, pp. 1–9.

Di Yuan is currently a Ph.D. student in the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China. She received B.Eng. degree in Electrical Engineering and Automatics from Tianjin University of Technology and Education, China, in 2016. Her research interests lie in medical image processing using deep learning methods and reinforcement learning.

Zhenghua Xu received a M.Phil. in Computer Science from The University of Melbourne, Australia, in 2012, and a D.Phil in computer Science from University of Oxford, United Kingdom, in 2018. From 2017 to 2018, he worked as a research associate at the Department of Computer Science, University of Oxford. He is now a professor at the Hebei University of Technology, China, and a awardee of "100 Talents Plan" of Hebei Province. He has published dozens of papers in top AI or database conferences, e.g., NeurIPS, AAAI, IJCAI, ICDE, etc. His current research focuses on deep learning, medical artificial intelligence, big data in health, and computer vision.

Biao Tian is currently a master student in the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China. He received B.Eng. degree in Electrical Engineering and Automatics from City College of Hebei University of Technology, China, in 2020. He research interests lie in medical image processing using deep learning methods.

Hening Wang is currently a master student in the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China. He received B.Eng. degree in Electrical Engineering and Automatics from Yanshan University, China, in 2022. He research interests lie in medical image processing using deep learning methods.

Yuefu Zhan is currently an Associate Chief Physician at Department of Radiology, Hainan Women and Children's Medical Center. He received a Doctor of Medicine degree from the West China Medical School, Sichuan University, China, in 2022, and a Master of Medicine degree from the Xiangya School of Medicine, Central South University, China, in 2010. His current research interests mainly focus on imaging diagnosis, minimally invasive intervention, and AI-based medical image analysis.

Thomas Lukasiewicz is a Professor of Computer Science at the Department of Computer Science, University of Oxford, UK, heading the Intelligent Systems Lab within the Artificial Intelligence and Machine Learning Theme. He currently holds an AXA Chair grant on "Explainable Artificial Intelligence in Healthcare" and a Turing Fellowship at the Alan Turing Institute, London, UK, which is the UK's National Institute for Data Science and Artificial Intelligence. He received the IJCAI-01 Distinguished Paper Award, the AIJ Prominent Paper Award 2013, the RuleML 2015 Best Paper Award, and the ACM PODS Alberto O. Mendelzon Test-of-Time Award 2019. He is a Fellow of the European Association for Artificial Intelligence (EurAI) since 2020. His research interests are especially in artificial intelligence and machine learning.