Neurocomputing 500 (2022) 177–190

Contents lists available at ScienceDirect

Neurocomputing

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

ω-net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution

Zhenghua Xu [a,b,*], Shijie Liu [a,b], Di Yuan [a,b,*], Lei Wang [a,b], Junyang Chen [c], Thomas Lukasiewicz [d], Zhigang Fu [e], Rui Zhang [f]

[a] State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China
[b] Tianjin Key Laboratory of Bioelectromagnetic Technology and Intelligent Health, Hebei University of Technology, China
[c] College of Computer Science and Software Engineering, Shenzhen University, China
[d] Department of Computer Science, University of Oxford, United Kingdom
[e] Department of Health Management Center, 983 Hospital of Joint Logistics Support Force, China
[f] www.ruizhang.info

ARTICLE INFO

Article history:
Received 28 January 2022
Revised 17 April 2022
Accepted 14 May 2022
Available online 18 May 2022
Communicated by Zidong Wang

Keywords:
Medical image segmentation
Dual supervision
Multi-dimensional self-attention
Diversely-connected multi-scale convolution

ABSTRACT

Although U-Net and its variants have achieved some great successes in medical image segmentation tasks, their segmentation performances for small objects are still unsatisfactory. Therefore, in this work, a new deep model, ω-Net, is proposed to achieve more accurate medical image segmentations. The advancements of ω-Net are mainly threefold: First, it incorporates an additional expansive path into U-Net to import an extra supervision signal and obtain a more effective and robust image segmentation by dual supervision. Then, a multi-dimensional self-attention mechanism is further developed to highlight salient features and suppress irrelevant ones consecutively in both spatial and channel dimensions. Finally, to reduce semantic disparity between the feature maps of the contracting and expansive paths, we further propose to integrate diversely-connected multi-scale convolution blocks into the skip connections, where several multi-scale convolutional operations are connected in both series and parallel. Extensive experimental results on three abdominal CT segmentation tasks show that (i) ω-Net greatly outperforms the state-of-the-art image segmentation methods in medical image segmentation tasks; (ii) the proposed three advancements are all effective and essential for ω-Net to achieve the superior performances; and (iii) the proposed multi-dimensional self-attention (resp., diversely-connected multi-scale convolution) is more effective than the state-of-the-art attention mechanisms (resp., multi-scale solutions) for medical image segmentations. The code will be released online after this paper is formally accepted.

1. Introduction

With the fast development of artificial intelligence, deep-learning-based medical image analysis technologies have been increasingly applied in clinical computer-aided diagnosis (CAD) [10,15]. Deep-learning-based medical image segmentation is one of the most important tasks in CAD [5], which aims to recognize and annotate the regions of interest (e.g., organs and lesions) with masks and/or outlines using deep models. U-Net is a widely exploited deep-learning-based medical image segmentation model, where a contracting path is utilized to extract deep features from the input images, an almost symmetric expansive path is used to achieve precise localization, and skip connections are adopted to remedy the information loss in convolutions [29].

Although U-Net has already achieved some great successes, its segmentation accuracy for small objects is still unsatisfactory. This is because the contracting path of U-Net will extract increasingly abstract or coarse feature maps layer by layer [11], so the features of some important small objects may become invisible or even be lost in the deep layers, making it difficult for U-Net to effectively learn features of small objects [9]. Skip connections are used in U-Net to remedy this problem by concatenating the deep and coarse features in the expansive path with the shallow and fine

* Corresponding author at: State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China.
E-mail addresses: zhenghua.xu@hebut.edu.cn (Z. Xu), yuandi.hn@gmail.com (D. Yuan).

https://doi.org/10.1016/j.neucom.2022.05.053
0925-2312/

features in the contracting path to enrich the feature information. However, this solution suffers from the following two shortcomings: (i) *Irrelevant information problem*: Concatenating features in the contracting path not only import the important missing information to the feature maps in the expansive path, but also introduce some irrelevant information that should have been filtered at the deeper layers of U-Net, which thus reversely affects the model's segmentation performances [28]. (ii) *Semantic disparity problem*: Since the concatenated feature maps are generated by different layers located at different depths of the deep network, there may exist semantic differences between them, so concatenating these feature maps directly may arguably be inappropriate and thus may weaken the segmentation accuracy [17,47]. Consequently, using solely skip connections is still not sufficient for U-Net to achieve accurate segmentations for small objects. Although the resulting segmenting errors may be relatively small, they are still unacceptable in practical medical image segmentation tasks, and may cause fatal consequences in clinical practice [47]. For example, when this model is used to delineate the target area of tumor radiotherapy, even a few tumor cells missed may cause the failure of radiotherapy and the recurrence of cancer. Therefore, the need of a more accurate deep model for medical image segmentation is compelling.

To overcome the irrelevant information problem, a variant of U-Net, Attention U-Net [28], has been proposed to utilize attention gates to assign the concatenated features with different weights to suppress irrelevant regions while highlighting the salient features that are useful for specific segmentation tasks. However, in Attention U-Net, the weights of concatenated features in a given layer are determined by the information of feature maps in the next layer; since the feature maps generated in deeper layers are more abstract (i.e., more likely to lose important features of minor objects), Attention U-Net may be able to highlight some salient features, but may also mistakenly suppress some important features of small objects that are missing in the next layer. U-Net++ [47], on the other hand, is another variant of U-Net, where the semantic disparity problem is alleviated by applying nested dense convolutions onto skip connections to generate feature maps that contain feature information with different scales. However, in U-Net++, the multi-scale feature information mainly comes from the more abstract feature maps generated in the deeper layers of the contracting path, which are usually more likely to lose some important information of small objects, so U-Net++ may be also inadequate for the small object segmentation tasks.

Consequently, in this work, we propose a dual supervised deep segmentation model, ω-*Net*, for more accurate medical image segmentation, where the irrelevant information problem and the semantic disparity problem are respectively remedied using a *multi-dimensional self-attention (MDSA)* mechanism and *diversely-connected multi-scale convolution (DC-MSC)* blocks. Generally, comparing to the conventional U-Net, the proposed ω-Net mainly has the following three improvements. First, in ω-Net, we incorporate an *additional expansive path* into U-Net to bring an additional learning loss (called *auxiliary loss*) for dual supervised segmentation, which can enhance the deep model's feature learning capability by obtaining features that are more effective and robust for image segmentation. Specifically, with the help of the additional expansive path, the segmentation learning in original expansive path takes into account not only the feature information from the contracting path, but also the intermediate segmentation information from the additional expansive path. Consequently, the deep model can generate more accurate segmentation results, because the final segmentation results can be seen as further refinements based on the coarse intermediate segmentation results generated in the additional expansive path.

Furthermore, the second advancement of ω-Net is to propose a novel *multi-dimensional self-attention (MDSA)* mechanism to remedy the irrelevant information problem using two consecutive self-attention modules, *dense spatial position attention (DSPA)* and *channel attention (CA)*, which respectively capture features' self-dependencies in the spatial and channel dimensions. Specifically, in DSPA, the importance of a position in a feature map is determined by its dependencies with all other positions in the same feature map; similarly, in CA, the importance of a channel is determined by its dependencies with all other channels within the same layer. Generally, MDSA has the following advantages: (i) The weights of features in MDSA are computed in both spatial and channel dimensions, making it describe the importance of features more comprehensively. (ii) The weights of features in MDSA are computed solely based on information sourced from themselves (i.e., self-attention [44]), so it will not encounter the same problem as Attention U-Net. (iii) More importantly, we notice that the computation of spatial attention in all the existing multi-dimensional attention works [8,38] is always very time and memory consuming, because the size of feature map is usually very large. So, in order to efficiently estimate the weights of features in the spatial dimension, MDSA applies a dilated convolution block in DSPA to convert the input feature map to a *dense feature matrix*, whose size is much smaller than the input feature map (i.e., representing the feature map in a much denser way), and then use the dense feature matrix, instead of using the input feature map directly, to estimate the spatial dependencies. Consequently, MDSA is more efficient than the existing multi-dimensional attention solutions, while achieving even better segmentation performances according to our experimental studies.

Finally, in order to alleviate the semantic disparity problem, we propose to further integrate *diversely-connected multi-scale convolution (DC-MSC)* blocks into the skip connections of ω-Net. DC-MSC utilizes convolution kernels with different sizes to generate feature maps that contain feature information of different scales, which thus reduces the semantic difference between the concatenated feature maps. Differently from the existing multi-scale methods [17,36,41,45], whose multi-scale pooling or convolution operations are all connected in parallel, the various-sized convolutional operations in DC-MSC are connected diversely in both series and parallel. We believe that the diverse connections of multi-scale convolution will enhance the utilization of the generated multi-scale feature maps, and help the deep model achieve better segmentation performances.

The contributions of this paper can be summarized as follows:

- We identify the existing two shortcomings of U-Net and propose a dual supervised deep model, ω-Net, to remedy these problems and achieve more accurate medical image segmentations.
- In ω-Net, an additional expansive path is first proposed to strengthen the deep segmentation model's feature learning capability based on dual supervision. Then, to overcome the irrelevant information problem, a multi-dimensional self-attention (MDSA) mechanism is further proposed to highlight the salient features and suppress the irrelevant ones using two consecutive self-attention modules to capture features' self-dependencies in both spatial and channel dimensions. Finally, we propose to integrate diversely-connected multi-scale convolution (DC-MSC) blocks into the skip connections to remedy the semantic disparity problem.
- Extensive experimental studies are conducted on three real-world abdominal CT segmentation datasets, and the results show the following: (i) The proposed ω-Net significantly outperforms the state-of-the-art image segmentation methods in the medical image segmentation tasks in terms of all metrics.

(ii) The proposed three advancements are all effective and essential for ω-Net to achieve the superior segmentation performances. (iii) The proposed multi-dimensional self-attention (resp., diversely-connected multi-scale convolution) is more effective than the state-of-the-art attention mechanisms (resp., multi-scale methods) in alleviating the irrelevant information (rep., semantic disparity) problem and achieving more accurate medical image segmentations.

2. Related Work

Medical image segmentation is the process of identifying and delineating the targeted objects (e.g., organs or lesions) in clinical images. Deep-learning-based methods have already been widely applied in medical image segmentation tasks. FCN is the first end-to-end image segmentation model using convolutional neural networks [25]; 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 [1,46]. To obtain more refined segmentations, 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 segmentations [29]. 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 segmenting brain tumor [2,6,42,43], liver [16,21,24], pancreas [28,36], and retinal vessels [33,34]. Despite achieving some successes, the performances of the existing U-Net based deep models are still unsatisfactory, especially for segmenting the small objects in medical images, so ω-Net is proposed in this work.

Improving the U-Net Architecture. The first contribution of ω-Net is to improve the architecture of U-Net by importing an additional expansive path. In recent years, there also exists some deep-learning-based segmentation methods that try to optimize the structure of U-Net for better segmentation performances [4,26,28,47]. To capture retinal vessels at various shapes and adaptively adjust the receptive fields according to the vessels' shapes, Jin et al. propose a deformable U-Net (DUNet), where deformable convolutions are introduced into the U-shape architecture to extract context information and enable precise localization [20]. To suppress the response of irrelevant background information and enhance the sensitivity of foreground information, Attention U-Net [28] is proposed to integrate attention gates into the expansive path to estimate the feature weights; while dual attention networks (DANet) [8] apply attention mechanisms in both spatial and channel dimensions. U-Net++ is introduced in [47] to apply nested dense convolutions onto skip connections to reduce the semantic gaps between the feature maps generated in the contracting and expansive paths; then U-Net3+ [16] further improves U-Net++ using full-scale skip connections and deep supervisions. Moreover, to better model the three-dimensional spatial relationship on 3D medical images, many 3D variants of U-Net have been proposed in recent works, e.g., 3D U-Net [4], V-Net [26], and nnU-Net [18]. Therefore, to evaluate the performances of ω-Net in medical image segmentation tasks, six state-of-the-art segmentation methods FCN [25], U-Net [29], Attention U-Net [28], DANet [8], U-Net++ [47], and U-Net 3+ [16] are selected as the baselines in our experiments.

Attention Mechanisms. The second improvement of ω-Net is to adopt a multi-dimensional self-attention (MDSA) mechanism to measure the importance of features on both spatial position and channel dimensions. Similarly, attention mechanisms [23,32] have also been utilized in some recent works to improve the performances of deep-learning-based image processing models [19,35,38,39]. In order to learn discriminative features and to address the specular reflection issue, Ni et al. propose a new attention module to capture global context and encode semantic dependencies to emphasize key semantic features [27]. Squeeze-and-excitation networks (SENets) are proposed in [14] to explicitly model inter-dependencies between channels and adaptively recalibrate channel-wise feature responses. As SENet may be sensitive to noise during average-pooling processes, Wang et al. propose to resolve this problem by using the weighted sum of the features at all positions to capture long-range dependencies for the global receptive field [37]. However, computing the weights of all the position is usually very time-consuming, to capture target objects with different scales, selective kernel networks (SKNet) uses pooling operations to compress the input feature maps and utilize softmax attention to fuse branches with different kernel sizes [22].

Similar to our work, there also exist some researches that propose to infer attention maps along two separate dimensions. Woo et al. propose a convolutional block attention module (CBAM) to sequentially infer channel and spatial attention maps, which are then multiplied to the input feature map for adaptive feature refinement [38]. Similarly, to integrate local features with their global dependencies adaptively, dual attention networks (DANets) are proposed to combine parallel position and channel attention (PPCA) modules with the traditional dilated FCN to respectively model the semantic inter-dependencies in spatial and channel dimensions [8]. However, the existing multi-dimensional attention works are usually very time and memory consuming, especially for high-resolution images, such as medical images. Differently, in this work, the proposed multi-dimensional self-attention (MDSA) mechanism uses dilated convolution blocks to generate dense feature matrices to decrease the size of the feature representations, which greatly reduces the time and memory consumption, while maintaining good performances in highlighting the salient features. Consequently, to show the effectiveness of MDSA in achieving better medical image segmentations, experimental studies are conducted in this work to also compare MDSA with the state-of-the-art attention mechanisms: squeeze-and-excitation attention (i.e., SOTA channel-wise attention) [14], selective kernel attention (i.e., SOTA spatial attention) [22], and parallel position and channel attention (i.e., SOTA multi-dimensional attention) [8].

Semantic Disparity and Multi-Scale Methods. The third advancement of ω-Net is to utilize diversely-connected multi-scale convolution blocks to resolve the semantic disparity problem. Recent years have witnessed some research works that also use multi-scale solutions to enhance the performances of convolutional deep models [7], where multi-scale mechanisms can be applied to either pooling or convolutional operations. A principled pooling strategy, spatial pyramid pooling (SPP), is proposed to enable the use of images with arbitrary sizes as the inputs of CNN-based deep models, which improves both CNN-based image classification and object detection methods in general [12]. In order to obtain an accurate segmentation performance in diverse scenes, Zhao et al. propose a pyramid scene parsing network (PSPNet), where a pyramid pooling module is adopted to generate and aggregate different-region-based context to better exploit global context information [45]. Besides multi-scale pooling, multi-scale convolution is also utilized to generate feature maps with different scales of details. The inception modules introduced by GoogLeNet are designed to process and aggregate visual information at different sizes of kernels to optimize the quality of feature learning [31]. Moreover, Yu et al. develop a new convolutional module, parallel dilated convolution (PDC), which uses dilated convolutions to

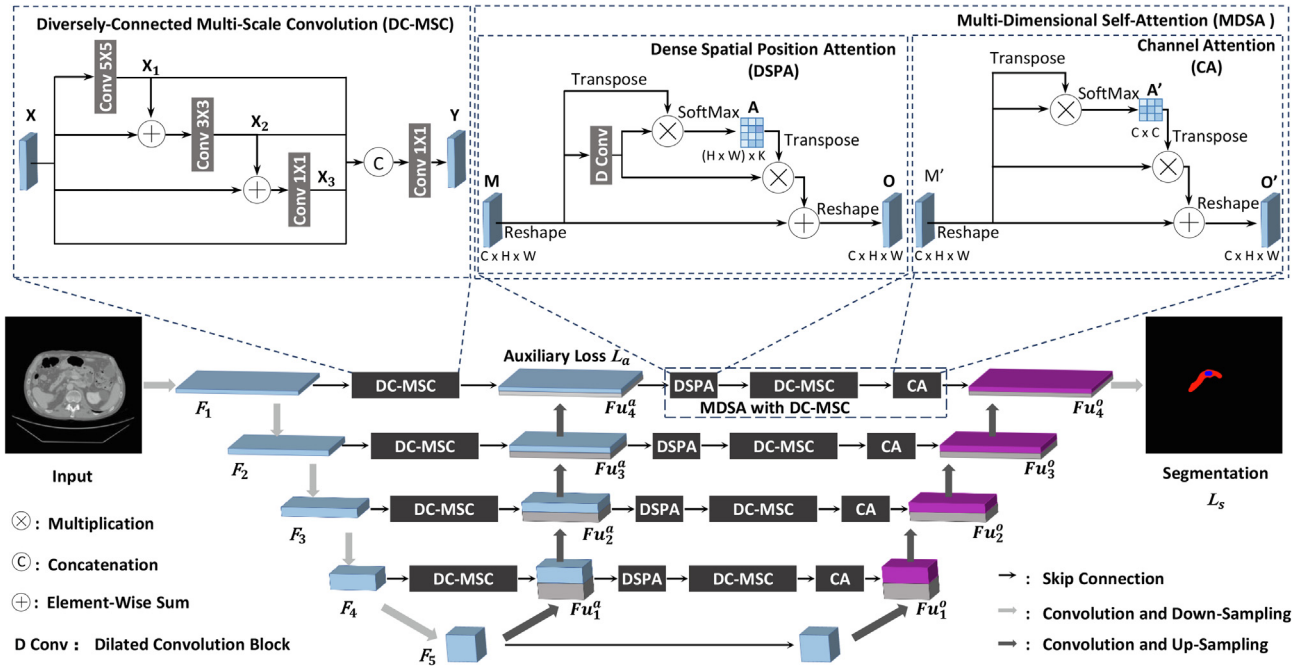


Fig. 1. Overall structure of ω-*Net*, where the blue blocks in additional expansive path are feature maps generated by the corresponding DC-MSC modules, and the purple blocks in original expansive path are feature maps generated by the corresponding MDSA with DC-MSC modules.

systematically aggregate multi-scale contextual information without losing resolution for dense segmentations [41].

Similar to this work, multi-scale convolutional operations are also utilized in [17] to alleviate the semantic disparity problem, where multi-scale convolution blocks that consist of two paralleled convolutional kernels are integrated into the skip connections of U-Net. Wang et al. also incorporate multi-scale parallel convolution modules into the skip connections, where the size of the multi-scale module is increased to three paralleled kernels [36]. We notice that the above related works all process the multi-scale pooling or convolution operations in parallel, in order to enhance the utilization of the generated multi-scale feature maps, diversely-connected multi-scale convolution (DC-MSC) blocks are proposed in this work, where the various-sized convolutional operations are connected in both series and parallel. To exhibit the superior performance of DC-MSC in medical image segmentations, experiments are also conducted to compare DC-MSC with three state-of-the-art multi-scale solutions: pyramid scene parsing network (i.e., SOTA multi-scale pooling mechanism) [45], parallel dilated convolution module (i.e., SOTA multi-scale convolution mechanism) [41], and parallel convolution module (i.e., SOTA multi-scale solution for the semantic disparity problem in U-Net) [36].

3. Dual Supervised Medical Image Segmentation with MDSA and DC-MSC

Fig. 1 shows the overall structure of ω-Net. Comparing to the conventional U-Net, ω-Net mainly consists of three additional advanced modules: additional expansive path, multi-dimensional self-attention (MDSA) mechanism, and diversely-connected multi-scale convolution (DC-MSC) blocks. Specifically, an additional expansive path is introduced to bring an additional learning loss, i.e., auxiliary loss, to strengthen the model's learning capability via dual supervision. Consequently, ω-Net can generate more accurate segmentation results using both the feature information from the contracting path and the intermediate segmentation information from the additional expansive path. Furthermore, an MDSA mechanism is proposed in ω-Net to resolve the irrelevant information problem by using two consecutive self-attention modules, dense spatial position attention (DSPA) and channel attention (CA), to capture the importance of features in both spatial and channel dimensions. To efficiently estimate the weights of features in the spatial dimension, MDSA applies a dilated convolution block in DSPA to convert the input feature map to a dense feature matrix with a smaller size, which is then used to estimate the spatial dependencies. Finally, DC-MSC blocks in ω-Net are used to remedy the semantic disparity problem using diversely-connected multi-scale convolution kernels, where the various-sized convolutional operations are connected in both series and parallel. Consequently, this enhances the utilization of the generated multi-scale feature maps and makes the final resulting feature maps of DC-MSC retain more comprehensive semantic information with different scales, which thus better reduces the semantic differences between the concatenated feature maps. Please note that although ω-Net aims at segmenting 2D medical images, it can be easily extended to 3D ω-Net to directly process 3D medical images using a way similar to extending U-Net to 3D U-Net [4], i.e., extending the convolution kernels and pooling operations from 2D to 3D, and keeping the number of channels unchanged.

3.1. Dual Supervised Segmentation with an Additional Expansive Path

The first improvement of ω-Net is to incorporate an additional expansive path into U-Net to achieve a more accurate medical image segmentation by dual supervision. Specifically, similarly to U-Net, ω-Net has only one contracting path; but, differently from U-Net, when the most abstract feature maps are obtained at the deepest layer of the contracting path, they are then sent to two expansive paths with similar structures. The expansive path that

is also included in U-Net is called *original expansive path*, and the newly added expansive path is called *additional expansive path*. In each layer of the additional expansive path, we concatenate the feature maps generated by the corresponding layer of the contracting path with the feature maps resulting from the transpose convolutional up-sampling operations in the last layer (actually, this process is the same as the skip connection operation in U-Net). Then, the concatenated feature maps in the additional expansive path are sent to its subsequent layer and also the corresponding layer of the original expansive path. Consequently, the up-sampled feature maps in each layer of the original expansive path are not only concatenated with the feature maps from the contracting path but also with the up-sampled feature maps from the additional expansive path. For example, as shown in Fig. 1, the purple block concatenated with Fu_4^o is the concatenation of F_1 (the corresponding feature map from the contracting path) and Fu_4^a (the corresponding feature map from the additional expansive path). Finally, with the help of the additional expansive path, ω-Net is learned by dual supervision, i.e., the deep model is trained using both a segmentation loss from the original expansive path and an auxiliary segmentation loss from the additional expansive path.

The advantages of introducing an additional expansive path to achieve dual supervision in ω-Net are twofold: On one hand, the additional auxiliary segmentation loss can help the feature learning process in the contracting path to obtain features that are more effective and robust for medical image segmentation; this is because, with various segmentation supervision losses, the features learned by the contracting path are now requested to adapt to two different segmentation optimization directions in two different expansive paths, which thus enhances the feature learning effectiveness and robustness of ω-Net. On the other hand, in ω-Net, the up-sampled feature maps in each layer of original expansive path are not only concatenated with the feature maps from the contracting path but also with the up-sampled feature maps from the additional expansive path; so the segmentation learning in the original expansive path takes into account not only the feature information from the contracting path, but also the intermediate segmentation information from the additional expansive path, which thus helps ω-Net generate more accurate segmentation results. The intuition is as follows, by treating the segmentation process in the additional expansive path as a coarse segmentation process, the final segmentation results generated in the original expansive path can now be seen as further refinements based on the coarse intermediate segmentation results generated in the additional expansive path, which are usually more accurate.

The formal definition of dual supervised segmentation with an additional expansive path can be written as follows. First, by denoting the input image as X, the output feature map in the first layer of the contracting path is formally

$$\mathbf{F}_1 = Conv^{1\times 64}(\mathbf{X}), \tag{1}$$

where $Conv^{1\times 64}$ means a convolutional operation whose number of input channels is 1 and the number of output channels is 64. Then, the output feature map at the i^{th} (where $i > 1$) layer of the contracting path is written as

$$\mathbf{F}_i = Pool\_Max(Conv\_2(\mathbf{F}_{i-1})), \tag{2}$$

where $Conv\_2(\cdot)$ represents two consecutive convolutional operations, and $Pool\_Max(\cdot)$ is a max-pooling operation.

Given F_d as the most abstract feature map generated in the last layer of the contracting path, where d is the number of layers in the contracting path, the feature map generated by the first transpose convolutional up-sampling in the additional expansive path can be formally written as

$$\mathbf{Fu}_1^a = De\_Conv(\mathbf{F}_d). \tag{3}$$

Then, with the skip-connection operation, the feature map generated by the j^{th} ($j > 1$) transpose convolutional up-sampling in the additional expansive path can be formally defined as

$$\mathbf{Fu}_j^a = De\_Conv(Concat(\mathbf{Fu}_{j-1}^a, \mathbf{F}_{d-j})), \tag{4}$$

where $Concat(\cdot)$ denotes the concatenation operation.

Similarly, the feature map generated by the first transpose convolutional up-sampling in the original expansive path can be formally defined as

$$\mathbf{Fu}_1^o = De\_Conv(\mathbf{F}_d). \tag{5}$$

Furthermore, the feature map generated by the j^{th} (where $j > 1$) transpose convolutional up-sampling in the original expansive path can be formally defined as

$$\mathbf{Fu}_j^o = De\_Conv(Concat(\mathbf{Fu}_{j-1}^o, Concat(\mathbf{Fu}_{j-1}^a, \mathbf{F}_{d-j}))), \tag{6}$$

where $Concat(\mathbf{Fu}_{j-1}^o, Concat(\mathbf{Fu}_{j-1}^a, \mathbf{F}_{d-j}))$ represents the operations that first concatenate $\mathbf{Fu}_{j-1}^a$ with $\mathbf{F}_{d-j}$ and then concatenate the resulting feature map with $\mathbf{Fu}_{j-1}^o$.

Finally, the auxiliary segmentation loss at the additional expansive path is formally defined as

$$\mathscr{L}_a = BCE(Conv^{64\times L}(\mathbf{Fu}_{d-1}^a), Mask), \tag{7}$$

where L is the number of channels of the given segmentation annotations, $Conv^{64\times L}(\cdot)$ is a convolutional operation whose number of input channels is 64, and the number of output channels is L, $BCE(\cdot)$ is the binary cross-entropy loss, and *Mask* denotes the corresponding segmentation annotations of input medical images. Similarly, the segmentation loss at the original expansive path can be formally written as

$$\mathscr{L}_s = BCE(\mathbf{Fu}_{d-1}^o, Mask). \tag{8}$$

Consequently, ω-Net is learned by a dual supervision loss that considers both the segmentation loss $\mathscr{L}_s$ and the auxiliary segmentation loss $\mathscr{L}_a$, formally,

$$\mathscr{L}_{dual} = \lambda\mathscr{L}_a + (1-\lambda)\mathscr{L}_s, \tag{9}$$

where λ is a hyperparameter that controls the relative importance of $\mathscr{L}_a$ and $\mathscr{L}_s$ in the dual supervision loss.

3.2. Multi-Dimensional Self-Attention

In ω-Net, we propose to utilize a multi-dimensional self-attention (MDSA) mechanism to remedy the irrelevant information problem using two consecutive self-attention [44] modules, dense spatial position attention (DSPA) and channel attention (CA), to capture features' self-dependencies in the spatial and channel dimensions, respectively. Generally, in DSPA, the importance of a position in a feature map is determined by its dependencies with all other positions in the same feature map; similarly, in CA, the importance of a channel is determined by its dependencies with all other channels. Consequently, by measuring the importance of features in two different dimensions, MDSA will describe the importance of features more comprehensively; furthermore, since CA and DSPA are both based on self-attention, MDSA will not encounter the same problem as Attention U-Net. We also note that, in Fig. 1, multi-dimensional self-attention (MDSA) blocks are added and only added into skip connections between the additional expansive path and the original expansive path; this is to make the segmentation model capable of conducting self-attention operations on all feature maps that are sent to original expansive path, while avoiding redundant computations.

Specifically, MDSA first uses a dense spatial position attention (DSPA) module to capture the spatial dependencies between each position in a feature map and all the positions in a dense feature matrix (called *dense positions*), where the dense feature matrix is generated by the corresponding feature map using a dilated convolution block (denoted D conv in Fig. 1). The reason of using the dilated convolution block to generate dense feature matrices instead of directly using the original feature map is to decrease the size of the feature representation, which not only greatly reduces the time and memory consumption but also helps MDSA to achieve even better segmentation performances (as shown in Section 4.7). Consequently, the feature value at a given position on the feature map is obtained by summarizing weighted feature values at all positions on the dense feature matrix, and the weights are based on the feature similarities between the given position on the feature map and the corresponding dense positions. With the help of DSPA, the features of small objects that are not salient on the feature map can now be enhanced using the salient features on the dense feature matrix that are highly similar to them, even if the salient features are extracted from regions that are far away from the small objects on the feature map.

The detailed operations and formal definitions of dense spatial position attention (DSPA) are as follows. In Fig. 1, the input feature map of DSPA, $\mathbf{M} \in \mathbb{R}^{C\times H\times W}$, generated at the j^{th} layer of the additional expansive path is obtained by concatenating the feature maps at the j^{th} layer of the additional expansive path, $\mathbf{Fu}_j^a$, and the ones from the corresponding layer of the contracting path, $\mathbf{F}_{d-j-1}$. Formally,

$$\mathbf{M} = Concat(\mathbf{Fu}_j^a, \mathbf{F}_{d-j-1}), \tag{10}$$

where d is the total number of layers of the contracting path.

We first reshape the input feature map $\mathbf{M}$ to $\mathbb{R}^{C\times N}$, where N is the total number of pixels in each channel. $\mathbf{M}$ is then sent into a dilated convolution block to generate a dense feature matrix $\mathbf{D} \in \mathbb{R}^{C\times K}$, where K is a hyperparameter representing the number of dense features in each channel of $\mathbf{D}$. Afterward, we perform a matrix multiplication between the transpose of $\mathbf{M}$ and $\mathbf{D}$, and use a softmax operation to calculate the dense spatial attention matrix $\mathbf{A} \in \mathbb{R}^{N\times K}$. Formally,

$$a_{j,i} = \frac{exp(\mathbf{D}_i \cdot \mathbf{M}_j)}{\sum_{i=1}^{N} exp(\mathbf{D}_i \cdot \mathbf{M}_j)}, \tag{11}$$

where $a_{j,i}$ is an element of the dense spatial attention matrix $\mathbf{A}$ located at the j^{th} raw and the i^{th} column, measuring the impact of the i^{th} feature of the dense feature matrix $\mathbf{D}$ on the j^{th} feature of the input feature map $\mathbf{M}$.

A matrix multiplication is conducted between $\mathbf{D}$ and the transpose of $\mathbf{A}$, whose result is then added to $\mathbf{M}$ using an element-wise sum operation. Finally, we reshape the summation result to get the final output feature map $\mathbf{O} \in \mathbb{R}^{C\times H\times W}$ of DSPA. Formally,

$$\mathbf{O}_j = \lambda_p \sum_{i=1}^{N} (a_{j,i}\mathbf{D}_i) + \mathbf{M}_j, \tag{12}$$

where λ_p is denoted as a strength coefficient, whose value is initialized to 0, and it is gradually learned to give appropriate importance to the spatial position attention map.

Similarly, a channel attention (CA) module is then used in MDSA to capture the channel dependencies between any two channel maps using also a self-attention procedure, where each channel map is updated by summarizing all weighted channel maps. Consequently, the inconspicuous features of small objects in a given channel map can now be enhanced by the salient features in other similar channel maps, which thus further remedies the information loss of small objects.

The detailed operations and formal definitions of channel attention (CA) are as follows. As shown in Fig. 1, given an input feature map of CA, $\mathbf{M}\prime \in \mathbb{R}^{C\times H\times W}$, we first reshape $\mathbf{M}\prime$ to $\mathbb{R}^{C\times N}$, and then perform a matrix multiplication between the transpose of $\mathbf{M}\prime$ and $\mathbf{M}\prime$, and use a softmax operation to calculate the channel attention matrix $\mathbf{A}\prime \in \mathbb{R}^{C\times C}$. Formally,

$$a\prime_{j,i} = \frac{exp(\mathbf{M}\prime_i \cdot \mathbf{M}\prime_j)}{\sum_{i=1}^{N} exp(\mathbf{M}\prime_i \cdot \mathbf{M}\prime_j)}, \tag{13}$$

where $a\prime_{j,i}$ is an element of the channel attention matrix $\mathbf{A}\prime$ located at the j^{th} raw and the i^{th} column, measuring the influence of the i^{th} channel on the j^{th} channel.

A matrix multiplication is conducted between $\mathbf{M}\prime$ and the transpose of $\mathbf{A}\prime$, whose result is then added to $\mathbf{M}\prime$ using an element-wise sum operation. Finally, we reshape the summation result to get the final output feature map $\mathbf{O}\prime \in \mathbb{R}^{C\times H\times W}$ of CA. Formally,

$$\mathbf{O}\prime_j = \lambda_c \sum_{i=1}^{N} (a\prime_{j,i}\mathbf{M}\prime_i) + \mathbf{M}\prime_j, \tag{14}$$

where λ_c controls the importance of the channel attention map over the input feature map. Similarly to λ_p, λ_c is also initially set to 0 and gradually learned in the model's training stage.

3.3. Diversely-Connected Multi-Scale Convolution

Another shortcoming of U-Net is that there exists a semantic disparity between the feature maps generated by the contracting path and the expansive path [17]. Therefore, the performances of U-Net-based segmentation models may be weakened if we concatenate these feature maps directly using skip connections. Actually, this problem also exists in the above proposed method: even if we have applied multi-dimensional self-attention on the skip connections of our segmentation model, MDSA only uses attention operations to highlight or suppress the information in the feature maps, but cannot effectively bridge the semantic gaps between the feature maps generated in the contracting path and the expansive path.

Therefore, in this work, a new multi-scale solution, diversely-connected multi-scale convolution (DC-MSC), is proposed to resolve the semantic disparity problem. Different from U-Net++, DC-MSC does not rely on deeper feature maps to get multi-scale features. Instead, it uses solely the feature maps generated in the given layer as the inputs and utilizes convolution kernels with different sizes to generate feature maps that contain feature information of different scales. The multi-scale feature maps are then fused to bridge the semantic gaps. Since all the multi-scale feature maps generated by the DC-MSC block at a given skip connection are based on the same source feature maps, and do not use any feature maps from deeper layer, the feature information of small objects can be retained in the generated multi-scale feature maps to a greater extent. Therefore, DC-MSC is a better choice for accurate small object segmentation.

Specifically, as shown in Fig. 1, by denoting the given input feature maps as $\mathbf{X}$, DC-MSC first sends $\mathbf{X}$ to a convolution block with the kernel size of 5×5 to get feature maps $\mathbf{X}_1$, whose size is the same as $\mathbf{X}$. Formally,

$$\mathbf{X}_1 = Conv\_5 \times 5(\mathbf{X}). \tag{15}$$

Then, $\mathbf{X}_1$ is added to $\mathbf{X}$ using an element-wise sum operation, and the summation result is processed by a convolution block with the kernel size of 3×3 to get feature maps $\mathbf{X}_2$. Formally,

Table 1
The information of datasets, where the average object sizes are in pixels.

Datasets	Source	Patients/	Training Set		Validation Set		Testing Set		Avg. Obj. Size	
		Samples	Samples	Images	Samples	Images	Samples	Images	Organ	Tumor
Kidney	KiTS19 [13]	210	147	31878	21	4781	42	8765	5568	2613
Pancreas	Decathlon [30]	281	197	18397	28	3088	56	5026	1589	640
Liver	Decathlon [30]	131	92	31982	13	9466	26	16566	17609	3689

$$\mathbf{X}_2 = Conv\_3\times 3(Sum(\mathbf{X}, \mathbf{X}_1)). \tag{16}$$

Similarly, $\mathbf{X}_2$ is added to $\mathbf{X}$ and the summation result is further processed by a 1×1 convolution block to obtain feature maps $\mathbf{X}_3$. Formally,

$$\mathbf{X}_3 = Conv\_1\times 1(Sum(\mathbf{X}, \mathbf{X}_2)). \tag{17}$$

Finally, we concatenate all feature maps $\mathbf{X}$, $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$, and apply another 1×1 convolution on the concatenation result to obtain the final output multi-scale feature maps $\mathbf{Y}$. Formally,

$$\mathbf{Y} = Conv\_1\times 1(Concat(\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)). \tag{18}$$

4. Experiments

4.1. Datasets

In order to evaluate the performances of our proposed ω-Net in diverse medical image segmentation tasks with different sizes of segmenting objects, we conduct extensive experiments on three computerized tomography (CT) image datasets (i.e., kidney, pancreas, and liver CT datasets), where each 3D CT sample is divided into different number of 2D CT images (varying from a few hundreds to more than a thousand), the images size is normalized to 512×512, about 70% of the samples in each dataset are selected as the training set, 10% as the validation set, and 20% as the testing set. The details of these datasets are presented in the following and their statistical information is shown in Table 1.

Kidney Data [13]: This is a public collection of segmented kidney CT images from 210 patients treated with partial or radical nephrectomy between 2010 and 2018. The average size of kidneys in this dataset is about 5568 pixels (about 2.1% of the whole image), while that of its tumors is about 2613 pixels (about 1.0% of the whole image); since the objects in this dataset are generally in medium size, it is selected in this work as a **medium object** segmenting task to show the performance of ω-Net in segmenting medium objects.

Pancreas Data [30]: This pancreas dataset is released by Memorial Sloan Kettering Cancer Center (New York, NY, USA), containing the CT samples and the corresponding segmentation annotations of 281 patients undergoing resection of pancreatic masses. The average size of pancreas in this dataset is about 1589 pixels (about 0.6% of the whole image), while that of its tumors is about 640 pixels (about 0.2% of the whole image); since the objects in this dataset are in very small size, it is selected in this work to show the superior performance of ω-Net in the challenging **small object** segmenting task.

Liver Data [30]: This public dataset contains the CT samples of liver and its tumors for 131 patients, where the corresponding semantic segmentation masks are provided by professional radiologists. The average size of livers in this dataset is about 17609 pixels (about 6.7% of the whole image), while that of its tumors is about 3689 pixels (about 1.4% of the whole image); since the objects in this dataset are in relatively large size, it is selected in this work as a **large object** segmenting task.

4.2. Baselines

In order to evaluate the performances of the proposed ω-Net, six state-of-art deep-learning-based image segmentation methods FCN [25], U-Net [29], Attention U-Net [28], DANet [8], U-Net++ [47], and U-Net 3+ [16] are selected as baselines. The reasons of selecting these six 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 ω-Net; (iii)**Attention U-Net** and **DANet** are the state-of-art solutions for the irrelevant information problem; and (iv) **U-Net++** and **U-Net 3+** are the state-of-art solutions for the semantic disparity problem.

4.3. Implementation Settings

Our experiments are implemented using the PyTorch framework[1] and run on two NVIDIA GeForce GTX 2080Ti GPUs. The implementation details of the proposed ω-Net are shown as follows. The contracting path of ω-Net consists of five layers, where each of the first four layers is built using two sequential 3×3 convolution operations and a 2×2 max-pooling operation, and the fifth layer contains only two 3×3 convolution operations. The structure of the additional expansive path is the same as that of the original expansive path, both of which consist of four layers with two sequential 3×3 convolution operations and a 2×2 transpose convolutional up-sampling operation in each layer. Furthermore, the numbers of kernels in the 1st to 5th layer of the contracting path are $64, 128, 256, 512$, and 1024, respectively, while the numbers of kernels in the 1st to 4th layer of the additional and original expansive paths are $512, 256, 128$, and 64, respectively. Finally, we add a 1×1 convolution block at the output layer to change the number of channels of the segmenting results from 64 to 2, depicting respectively the segmentation masks of the organs and tumor lesions.

ω-Net and all the baselines are trained using the Adam optimizer with a mini-batch size of 2, where the weight decay parameter in Adam is set to 0.00015. The learning rate is initialized as 0.0001 and decays with the rise of the number of training epochs; specifically, we first multiply the learning rate by a decay factor of 0.6 at the end of the 5th epoch and then repeatedly multiply it by 0.6 every three epochs during training. Finally, we have conducted a grid search to investigate the effect of varying two hyperparameters λ (the training loss coefficient defined in Eq. (9), which controls the weight of auxiliary loss and segmentation loss) and K (the size of the dense feature metrics in the DSPA module) on the segmentation performance of ω-Net, and set $\lambda = 0.15$ and $K = 256$; detailed information on the grid search is presented in Section 4.9. Dropout techniques [40] can also be applied to prevent over-fitting.

4.4. Evaluation Metrics

To evaluate the segmentation performances of our proposed ω-Net and the state-of-art baselines, three widely used segmentation

[1] https://pytorch.org/

Table 2
The results of our proposed ω-Net and the state-of-the-art medical image segmentation baselines on three abdominal CT segmentation datasets with different sizes of segmenting objects, where the best and the second best results are bold and underlined, respectively.

Architecture	Kidney (Medium)			Pancreas (Small)			Liver (Large)		
	DSC	PPV	Sensi	DSC	PPV	Sensi	DSC	PPV	Sensi
FCN [25]	0.7869	0.7947	0.7717	0.5162	0.6068	0.5047	0.8024	0.7905	0.7992
U-Net [29]	0.8561	0.8666	0.8309	0.5870	0.6523	0.5612	0.8535	0.8669	0.8416
U-Net++ [47]	0.8879	0.9126	0.8725	0.6256	0.6902	0.6075	0.8992	0.9101	0.8851
Attention U-Net [28]	0.8864	0.9001	0.8714	0.6355	0.6924	0.6150	0.8840	0.8971	0.8753
DANet [8]	0.8819	0.8839	0.8837	0.6316	0.6688	0.6231	0.9031	0.9092	0.8981
U-Net 3+ [16]	0.8919	0.9005	0.8863	0.6329	0.6970	0.5742	0.9021	0.9172	0.8789
ω-Net (ours)	**0.9047**	**0.9339**	**0.8964**	**0.6495**	**0.7178**	**0.6349**	**0.9111**	**0.9230**	**0.9107**
Improvement	0.0128	0.0213	0.0101	0.0140	0.0208	0.0118	0.0080	0.0058	0.0126

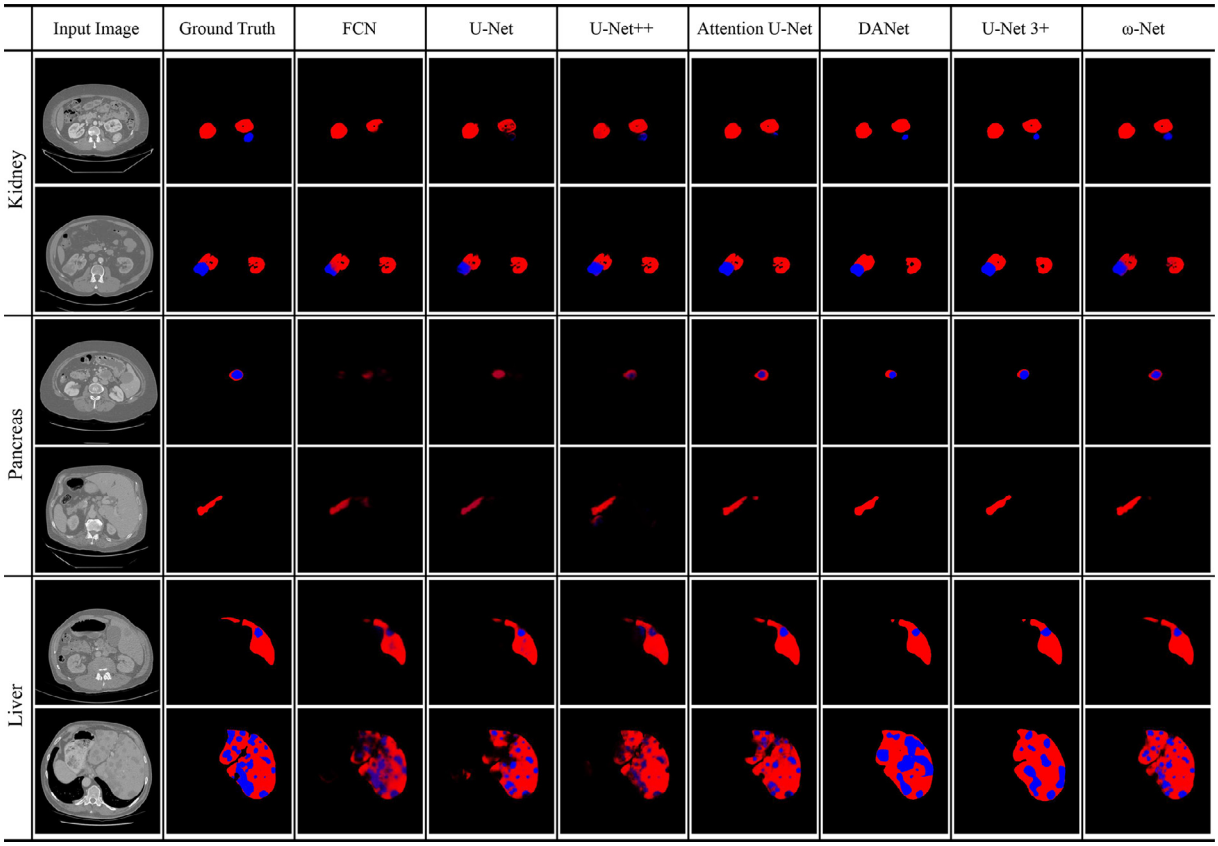


Fig. 2. Examples of visualized segmentation results of our proposed ω-Net and the baselines on three CT datasets, where red masks are for organs, and blue masks are for tumors.

evaluation metrics, *positive predictive value (PPV)*, *sensitivity (Sensi)*, and *dice similarity coefficient (DSC)*, are adopted. The formal definitions of DSC, PPV, and Sensi are as follows.

$$PPV = \frac{TP}{TP+FP}, \qquad Sensi = \frac{TP}{TP+FN},$$
$$DSC = \frac{2\times PPV\times Sensi}{PPV+Sensi} = \frac{2TP}{2TP+FP+FN},$$

where *TP*, i.e., true positive, is the number of positive pixels (i.e., pixels within the annotated organ or tumor areas) that are correctly classified in the segmenting results; *FP*, i.e., false positive, is the number of negative pixels (i.e., pixels within annotated background areas) that are incorrectly classified as positive pixels; *FN*, i.e., false negative, is the number of positive pixels that are incorrectly classified as negative pixels. Specifically, positive predictive value (PPV), also known as precision [3], represents the proportion of positive pixels that are correctly segmented to all the pixels that are classified as positive in the segmenting results. Sensitivity, also known as recall [3], is the proportion of positive pixels that are correctly segmented to all the pixels that are annotated as positive in the ground truths. Dice similarity coefficient (DSC), also known as F1-score [3], is the harmonic mean of PPV and Sensi, which thus can evaluate the model's performances more comprehensively from the perspectives of both PPV and Sensi.

4.5. Main Results

The experimental results of ω-Net and six state-of-the-art medical image segmentation baselines on three abdominal CT segmentation datasets with different sizes of segmenting objects are shown in Table 2, while six examples of visualized segmentation results are shown in Fig. 2.

As shown in Table 2, Attention U-Net and DANet are generally better than FCN and U-Net on all datasets in terms of all metrics. This is because the attention mechanism can assign the

Table 3
Ablation studies on three datasets with different sizes of segmenting objects, where the best results are bold.

Architecture	Kidney (Medium)			Pancreas (Small)			Liver (Large)		
	DSC	PPV	Sensi	DSC	PPV	Sensi	DSC	PPV	Sensi
U-Net [29]	0.8561	0.8666	0.8309	0.5870	0.6523	0.5612	0.8535	0.8669	0.8416
U-Net with AEP	0.8614	0.8762	0.8459	0.5954	0.6771	0.5817	0.8601	0.8774	0.8591
U-Net with MDSA	0.8697	0.8906	0.8525	0.5993	0.6776	0.5872	0.8719	0.8767	0.8639
U-Net with DC-MSC	0.8756	0.9009	0.8546	0.6027	0.6789	0.5893	0.8823	0.8879	0.8750
U-Net with AEP + MDSA	0.8784	0.8995	0.8746	0.6062	0.6808	0.5923	0.8772	0.8936	0.8819
U-Net with AEP + DC-MSC	0.8902	0.9167	0.8818	0.6107	0.6905	0.6076	0.8993	0.9095	0.8893
ω-Net (ours)	**0.9047**	**0.9339**	**0.8964**	**0.6495**	**0.7178**	**0.6349**	**0.9111**	**0.9230**	**0.9107**

concatenated features with different weights to suppress irrelevant regions, while highlighting the salient features that are useful for specific segmentation tasks. This proves the existence of the irrelevant information problem and also proves that deep models' segmentation performances can be improved by solving the irrelevant information problem using attention mechanisms. Then, we observe that U-Net++ and U-Net 3+ consistently outperform FCN and U-Net, which is because U-Net++ and U-Net 3+ alleviate the semantic disparity problem by respectively applying nested dense and full-scale skip connections to generate feature maps that contain feature information with different scales. This thus proves that the segmentation performance of models can be improved by alleviating the semantic disparity problem through multi-scale solutions. Finally, we find that our proposed ω-Net generally outperforms all the baselines on all three datasets, which proves that ω-Net achieves better performances than the state-of-art image segmentation solutions in diverse medical image segmentation tasks. The reasons of superior performances of ω-Net are as follows: (i) ω-Net utilizes a multi-dimensional self-attention (MDSA) mechanism and diversely-connected multi-scale convolution (DC-MSC) blocks to resolve both the irrelevant information and the semantic disparity problems, (ii) our additional experiments in Section 4.7 (resp., 4.8) proves that the proposed MDSA mechanism (resp., DC-MSC blocks) can achieve a better improvement in medical image segmentation than the state-of-the-art attention mechanisms (resp., multi-scale solutions), and (iii) ω-Net additionally introduces an additional expansive path to strengthen the model's learning capability.

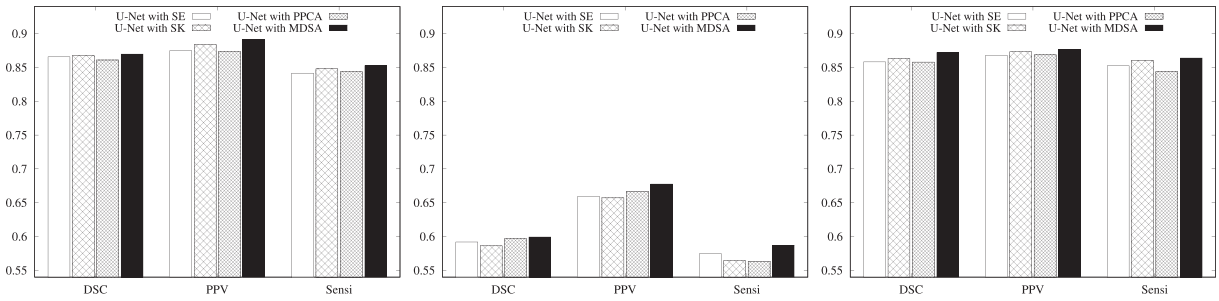
In addition, we also find in Table 2 that comparing to the performances of the best baselines, ω-Net generally achieves the highest performance improvements on the pancreas dataset, while the performance improvements are lowest on the liver dataset. This is because the average sizes of segmentation objects in the liver dataset are much larger than those in the pancreas dataset (about 10 times larger for organ and 5 times larger for tumor as shown in Table 2), making the pancreas segmentation tasks suffer from more severe irrelevant information and semantic disparity problems than the liver segmentation tasks. Consequently, this demonstrates the following conclusions: (i) the smaller the segmenting objects, the more severe the irrelevant information and semantic disparity problems are; (ii) ω-Net achieves superior medical image segmentation performances by resolving the irrelevant information and the semantic disparity problems using MDSA and DC-MSC.

Fig. 2 shows the visualized segmentation results of ω-Net and the baselines on six examples from three datasets. Specifically, the kidney images (at the first two rows) show that: (i) the segmentation results of FCN and U-Net are very incorrect in both kidney and tumors and sometimes even missing the tumor objects; (ii) those of U-Net++, Attention U-Net, DANet, and U-Net 3+ are relatively better for segmenting a kidney, but their segmentation results for the tumor objects are still poor; and (iii) the segmentation performance of ω-Net is much better than the baselines, its segmentation results for the small tumor objects are very close to ground truths. Similarly, we have the following observations for the pancreas images. (i) FCN and U-Net can neither correctly recognize and segment the pancreas nor the tumor; (ii) U-Net++, Attention U-Net, DANet, and U-Net 3+ are better but their performances in segmenting the edge areas of pancreas and tumor are not satisfactory; and (iii) the segmentation results of the proposed ω-Net are best among all models. Similar observations are also found for the liver images, especially for the case at the last row of Fig. 2, where ω-Net is the only model that correctly segments most of the small tumor lesions within the given liver image. Therefore, these visualized examples greatly demonstrate again that by the proposed MDSA mechanism, DC-MSC blocks, and the additional expansive path, ω-Net remedies the drawbacks of the existing deep segmentation models, and achieves much better performances in medical image segmentation tasks, especially for small objects.

4.6. Ablation Study

To show the effectiveness and necessity of the proposed three advanced modules in ω-Net, ablation studies are further conducted, where several intermediate models that only use one or two advanced modules are introduced and evaluated. Specifically, the intermediate models are as follows: (i) **U-Net with AEP** is a model that adds the additional expansive path into U-Net; (ii) **U-Net with MDSA** is constructed by adding the multi-dimensional self-attention (MDSA) module onto the skip connections of U-Net; (iii) **U-Net with DC-MSC** is obtained by adding the diversely-connected multi-scale convolution (DC-MSC) module onto the skip connections of U-Net; (iv) **U-Net with AEP + MDSA** integrates both the proposed additional expansive path and the multi-dimensional self-attention module with U-Net; and (v) **U-Net with AEP + DC-MSC** incorporates both the proposed additional expansive path and the diversely-connected multi-scale convolution module into U-Net.

In Table 3, all five intermediate models outperform U-Net over all three databases in terms of all metrics, which proves that the proposed advanced modules are all effective to improve the performance of U-Net in medical image segmentation tasks. Specifically, we first compare the results of U-Net and U-Net with AEP, where U-Net with AEP outperforms U-Net on all datasets in terms of all metrics. This is because the additional expansive path can enhance the deep model's feature learning capability by obtaining features that are more effective and robust for image segmentation. This thus proves that it is effective to improve the segmentation model's performance through adding an additional expansive path in ω-Net to achieve dual supervision. Then, it is observed that U-Net with MDSA and U-Net with DC-MSC consistently outperform U-Net in Table 3. This is because U-Net with MDSA (resp., U-Net with DC-MSC) resolves the irrelevant information problem (resp., semantic disparity problem) using a novel attention mechanism (multi-scale solution). Therefore, this proves the effectiveness of the multi-dimensional self-attention mechanism and diversely-connected multi-scale convolution blocks in medical image segmentation tasks. Furthermore, we notice that U-Net with



(a) Kidney (b) Pancreas (c) Liver

Fig. 3. The results of incorporating multi-dimensional self-attention and the state-of-the-art attention mechanisms into U-Net.

Table 4
The numerical results of incorporating multi-dimensional self-attention and the state-of-the-art attention mechanisms into U-Net, where the best results are bold.

Architecture	Kidney			Pancreas			Liver		
	DSC	PPV	Sensi	DSC	PPV	Sensi	DSC	PPV	Sensi
U-Net with SE [14]	0.8656	0.8748	0.8415	0.5919	0.6593	0.5747	0.8582	0.8677	0.8524
U-Net with SK [22]	0.8673	0.8837	0.8479	0.5862	0.6571	0.5642	0.8632	0.8731	0.8603
U-Net with PPCA [8]	0.8609	0.8739	0.8438	0.5971	0.6669	0.5635	0.8579	0.8685	0.8439
U-Net with MDSA	**0.8697**	**0.8906**	**0.8525**	**0.5993**	**0.6776**	**0.5872**	**0.8719**	**0.8767**	**0.8639**

AEP + MDSA is always better than U-Net with AEP and U-Net with MDSA. This is because the two advanced modules, AEP and MDSA, improve the segmentation performance of U-Net by tackling different problems, i.e., AEP is used to get more effective and robust features, while MDSA is used to resolve the irrelevant information problem. This thus proves that it is reasonable to incorporate both AEP and MDSA into the U-Net to achieve more accurate segmentation results. Similar observations and conclusions are also obtained by comparing U-Net with AEP + DC-MSC to U-Net with AEP and U-Net with DC-MSC. Finally, we find that ω-Net constantly achieves much better performances than U-Net with AEP + MDSA and U-Net with AEP + DC-MSC. This is because the three advanced modules target at resolving different problems, and can complement each other to better improve the deep model's segmentation performance. Therefore, the above observations demonstrate that the proposed three advanced modules are all effective and essential for ω-Net to achieve the superior medical image segmentation performances.

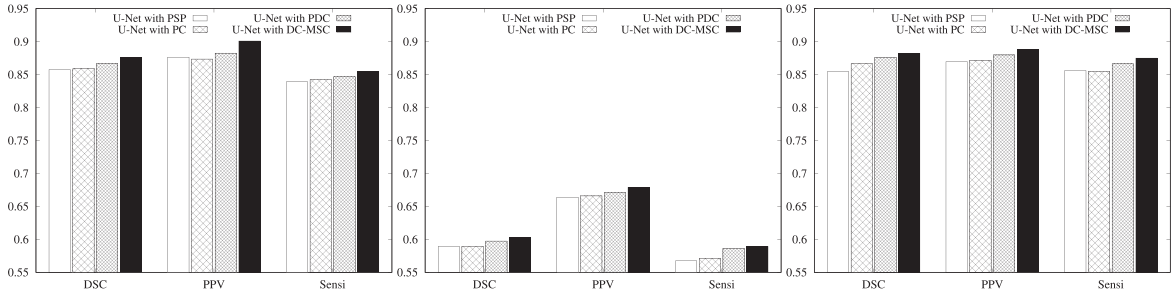
4.7. Multi-Dimensional Self-Attention vs. the State-of-the-art Attention Mechanisms

Further experiments are conducted to compare our proposed multi-dimensional self-attention (MDSA) module with the state-of-the-art attention mechanisms, namely, squeeze-and-excitation (SE) attention [14] (the state-of-the-art channel-wise attention), selective kernel (SK) attention [22] (the state-of-the-art spatial attention), and parallel position and channel attention (PPCA)[8] (the state-of-the-art multi-dimensional attention), where the different attention blocks are respectively incorporated with U-Net to show their different capabilities in enhancing U-Net's performances in medical image segmentation tasks. The corresponding experimental results are depicted in Fig. 3 and recorded in Table 4. The results show that incorporating U-Net with our proposed MDSA module can achieve much better performance improvements than using the state-of-the-art channel-wise attention (SE), spatial attention (SK), and multi-dimensional attention (PPCA) mechanisms, in terms of all metrics on all three datasets. This finding thus proves that MDSA is a better choice for medical image segmentation tasks than the state-of-the-art attention mechanisms.

4.8. Diversely-Connected Multi-Scale Convolution vs. the State-of-the-art Multi-Scale Solutions

Similarly, to investigate the influence of different multi-scale solutions on the performance of medical image segmentation, experiments are further conducted to compare the diversely-connected multi-scale convolution (DC-MSC) module with the state-of-the-art multi-scale solutions, namely, pyramid scene parsing (PSP) module [45], parallel convolution (PC) module [36], and parallel dilated convolution (PDC) module [41]), where the multi-scale modules are integrated into the skip connection of U-Net.

Generally, as shown in both Fig. 4 and Table 5, the model of combining U-Net with our proposed DC-MSC module (denoted U-Net with DC-MSC) outperforms the models of combining U-Net with the state-of-the-art multi-scale solutions in terms of all metrics on all three datasets, which proves that the proposed DC-MSC module can achieve a more accurate medical image segmentation than the state-of-art multi-scale solutions. Specifically, the performances of all the multi-scale convolution based models (U-Net with PC, U-Net with PDC, and U-Net with DC-MSC) are better than that of the multi-scale pooling based model, U-Net with PSP. This shows that applying multi-scale strategies on the convolution operations may be better than on the pooling operations, so in our DC-MSC, we apply the diversely-connected multi-scale strategy on convolution operations. Furthermore, we note that U-Net with DC-MSC is much better than U-Net with PC and U-Net with PDC in all the cases, which is because DC-MSC connects the various-sized convolutional operations diversely in both series and parallel, while PC and PDC only connect them in parallel, making DC-MSC capable of utilizing the generated multi-scale feature maps more comprehensively. In summary, these findings clearly demonstrate the effectiveness and reasonableness of the proposed DC-MSC in achieving better medical image segmentation performances than the state-of-the-art multi-scale solutions.



(a) Kidney (b) Pancreas (c) Liver

Fig. 4. The results of incorporating diversely-connected multi-scale convolution and the state-of-the-art multi-scale solutions into U-Net.

Table 5
The numerical results of incorporating diversely-connected multi-scale convolution and the state-of-the-art multi-scale solutions into U-Net, where the best results are bold.

Architecture	Kidney			Pancreas			Liver		
	DSC	PPV	Sensi	DSC	PPV	Sensi	DSC	PPV	Sensi
U-Net with PSP [45]	0.8575	0.8764	0.8391	0.5893	0.6639	0.5674	0.8546	0.8695	0.8561
U-Net with PC [36]	0.8587	0.8735	0.8422	0.5889	0.6658	0.5715	0.8663	0.8715	0.8552
U-Net with PDC [41]	0.8671	0.8825	0.8467	0.5974	0.6706	0.5864	0.8759	0.8797	0.8663
U-Net with DC-MSC	**0.8756**	**0.9009**	**0.8546**	**0.6027**	**0.6789**	**0.5893**	**0.8823**	**0.8879**	**0.8750**

4.9. Effect of Varying Hyper-Parameters K and λ

As shown in Eq. 9, a coefficient λ is used in the final segmentation loss function of ω-Net to control the relative importance of the segmentation loss $\mathscr{L}_s$ and the auxiliary segmentation loss $\mathscr{L}_a$. Therefore, the value of λ will greatly influence the model's training quality and also the final segmentation performances. Similarly, as a hyperparameter, K is introduced into the dilated convolution block of multi-dimensional self-attention (MDSA) module to decide the number of dense features used in the dense feature matrix D, whose value thus greatly affects the effectiveness of MDSA as well as the model's final segmentation performances. Consequently, experiments are conducted to investigate the effect of varying the hyperparameters λ and K on the model's training quality in terms of validation losses.

Since the segmentation loss $\mathscr{L}_s$ is obviously more important than the auxiliary segmentation loss $\mathscr{L}_a$, the value of λ is selected incrementally from 0.05 to 0.3 with a step of 0.05. To enhance the efficiency of computing using GPU, the value of K is set to 2^n, where the value of n is searched in $\{4, 5, 6, 7, 8, 9\}$. Generally, in Fig. 5, we observe that ω-Net obtains the lowest validation loss when $K = 256$ and λ=0.15, which are thus used as the final selected values.

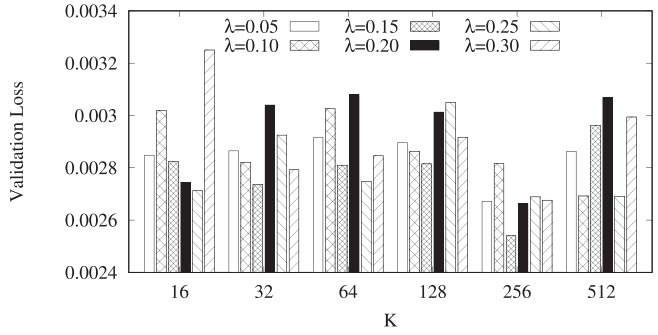


Fig. 5. Validation losses of ω-Net under different settings of hyper-parameters K and λ.

Moreover, the results in Fig. 5 show as another important finding that the validation losses of ω-Net fluctuate with the increase of the value of K. This finding thus further proves the **reasonableness of using a dense feature matrix**, instead of using the input feature map directly, in the MDSA modules of ω-Net: Since the accuracy of the spatial dependency estimation in MDSA is not guaranteed to increase with the rise of the size of the dense feature matrix D, it is reasonable to say that using a dense feature matrix whose size is much smaller than the input feature map in MDSA does not necessarily weaken the accuracy of the feature weight estimation and the model's feature learning capability, while it is guaranteed to greatly enhance the model's training efficiency. So, with a proper tuning of the hyperparameter K, using the dense feature matrix can greatly increase not only the efficiency but also the effectiveness of ω-Net.

4.10. Training and Inference Efficiency

Table 6 shows the training time–cost training time–cost (in hours per epoch) and inference efficiency (in images per second) of ω-Net and the state-of-the-art baselines. Generally, we notice that the more complicated the model is, the lower its training and inference efficiency. Specifically, we have the following observations: (i) By applying attention mechanisms to resolve the irrelevant information problem, Attention U-Net and DANet have lower training and inference efficiency. (ii) Using multi-scale solutions to overcome the semantic disparity problem inevitably enhances the training time–cost and reduces the inference efficiency of U-Net++ and U-Net 3+. (iii) Due to the use of additional expansive path for dual supervision and using the more powerful but also more complicated modules, MDSA and DC-MSC, to resolve the irrelevant information and semantic disparity problems, ω-Net inevitably has higher time–cost and lower inference efficiency than the baselines. However, the training and inference efficiency of ω-Net is still close to that of U-Net 3+, so considering the increasing computing capability of current facilities, sacrificing a limited extent of efficiency for better accuracy is acceptable for medical image segmentation tasks.

Table 6
The training time–cost (in hours per epoch) and inference efficiency (in images per second) of ω-Net and the state-of-the-art baselines.

	FCN	U-Net	Attention U-Net	U-Net++	DANet	U-Net 3+	ω-Net
Kidney	0.2373	0.2177	0.3834	1.1957	1.5222	3.0690	3.7063
Pancreas	0.1001	0.1012	0.1783	0.6074	0.6653	1.2577	1.6846
Liver	0.1916	0.1893	0.3195	1.1244	1.3896	2.8306	3.2257
Inference	20.52	19.57	16.75	8.71	19.05	6.66	4.22

5. Discussion and Future Work

5.1. Social Impact of ω-Net

The proposed ω-Net can be widely used in a lot of clinical scenarios, where the work of segmenting medical images is needed to effectively reduce the workload of doctors and improve the efficiency and accuracy of medical image segmentation. We take radiotherapy for cancer as an example, where doctors need to accurately delineate the outline of the tumor area on the patient's 3D CT images as the radiotherapy target area. However, each 3D CT is composed of hundreds of slices, and will take an experienced doctor several hours to annotate them one by one. Moreover, since the edge of the tumor is uneven and very difficult to delineate, to ensure the accuracy and comprehensiveness of labeling, it is usually necessary for multiple doctors to label the same image independently, and then gather them together as the final results. Consequently, the whole image segmentation process is very time-consuming and laborious; this not only greatly consumes the medical social resources (e.g., the time of experienced doctors), but may also bring long waiting times for the patient and delay the treatment. By applying our proposed automatic segmentation solution, ω-Net, in such clinical practices, the model can generate the draft segmentation results automatically in seconds, which can then be sent to experienced doctors for fine-tuning. This thus greatly reduces the workload of doctors, and saves both time and money for patients.

5.2. Limitation and Future Work

Despite achieving generally a superior performance in medical image segmentation tasks, we also observe in the experimental results that the performance of all segmentation models, including ω-Net, on the Pancreas dataset are much worse than those on the Kidney and Liver datasets. This is because the shape and appearance of the pancreas in medical images are much more various than those of the kidney and liver, so it is more difficult for the deep model to learn its morphological features. Therefore, it is an interesting future work to further improve the feature learning modules in ω-Net to resolve this problem and make ω-Net more applicable in the segmentation task of the pancreas and other morphologically various objects. In addition, in the future, it will be interesting to also conduct more experiments to investigate the performances of ω-Net in more diverse medical image segmentation tasks with different types of medical images, e.g., MRI, PET, X-ray, etc.

6. Conclusion

In this work, we identified two shortcomings of U-Net, namely, the irrelevant information problem and the semantic disparity problem, and proposed a novel dual supervised medical image segmentation model, called ω-Net, to remedy these problems and achieve a more accurate medical image segmentation using a multi-dimensional self-attention (MDSA) mechanism and diversely-connected multi-scale convolution (DC-MSC) blocks. Specifically, the technical contributions of ω-Net are threefold: We first integrate an additional expansive path into U-Net to introduce an extra supervision signal, called auxiliary loss, to obtain a more effective and robust image segmentation by the dual supervision. Then, the MDSA mechanism is proposed in ω-Net to resolve the irrelevant information problem by using two consecutive self-attention modules to capture features' importance in both spatial and channel dimensions. Finally, to remedy the semantic disparity problem, DC-MSC blocks are proposed and integrated into the skip connections of ω-Net, where several multi-scale convolutional operations are diversely connected in both series and parallel to utilize the generated multi-scale feature maps more comprehensively. Extensive experimental studies are conducted on three real-world medical image segmentation datasets, and the results show that the proposed ω-Net can significantly outperform the state-of-the-art image segmentation solutions in medical image segmentation tasks in terms of all metrics, and the additional expensive path, MDSA, and DC-MSC are all effective and essential for ω-Net to achieve the superior segmentation performance.

CRediT authorship contribution statement

Zhenghua Xu: Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Resources, Supervision, Funding acquisition. **Shijie Liu:** Investigation, Visualization, Software, Writing - review & editing. **Di Yuan:** Methodology, Software, Formal analysis, Writing - review & editing, Project administration. **Lei Wang:** Supervision, Resources, Writing - review & editing. **Junyang Chen:** Conceptualization, Writing - review & editing. **Thomas Lukasiewicz:** Conceptualization, Supervision, Writing - review & editing. **Zhigang Fu:** Data curation, Writing - review & editing. **Rui Zhang:** Resources, Writing - review & editing.

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 grant 61906063, by the Natural Science Foundation of Hebei Province, China, under the grant F2021202064, by the Natural Science Foundation of Tianjin City, China, under the grant 19JCQNJC00400, by the "100 Talents Plan" of Hebei Province, China, under the grant E2019050017, and by the Yuanguang Scholar Fund of Hebei University of Technology, China. This work was also partially supported by the AXA Research Fund.

References

[1] A. Ben-Cohen, I. Diamant, E. Klang, M. Amitai, H. Greenspan, Fully Convolutional Network for Liver Segmentation and Lesions Detection, in: Deep Learning and Data Labeling for Medical Applications, 2016, pp. 77–85.

[2] W. Chen, B. Liu, S. Peng, J. Sun, X. Qiao, S3d-unet: Separable 3D U-Net for Brain Tumor Segmentation, in: International MICCAI Brainlesion Workshop, 2018, pp. 358–368.
[3] N. Chinchor, B.M. Sundheim, Muc-5 evaluation metrics, in: Proceedings of The Fifth Message Understanding Conference, 1993, pp. 69–78.
[4] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning Dense Volumetric Segmentation From Sparse Annotation, in: Proceedings of The International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 424–432.
[5] 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.
[6] H. Dong, G. Yang, F. Liu, Y. Mo, Y. Guo, Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks, in: Proceedings of The Annual Conference on Medical Image Understanding and Analysis, 2017, pp. 506–517.
[7] R. Du, J. Xie, Z. Ma, D. Chang, Y.Z. Song, J. Guo, Progressive learning of category-consistent multi-granularity features for fine-grained visual classification, IEEE Transactions on Pattern Analysis and Machine Intelligence Early Access (2021) 1.
[8] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual Attention Network for Scene Segmentation, in: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[9] 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.
[10] 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, Computer Methods and Programs in Biomedicine 158 (2018) 113–122.
[11] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, S. Hikosaka, Effective Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Imagery, in: Proceedings of The IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 1442–1450.
[12] K. He, X. Zhang, S. Ren, J. Sun, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (2015) 1904–1916.
[13] Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al., 2019. The Kits19 Challenge Data: 300 Kidney Tumor Cases With Clinical Context, CT Semantic Segmentations, and Surgical Outcomes. arXiv preprint arXiv:1904.00445.
[14] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, in: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[15] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, Q. Sun, Deep Learning for Image-Based Cancer Detection and Diagnosis- A Survey, Pattern Recognition 83 (2018) 134–149.
[16] 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 UNet for Medical Image Segmentation, in: Proceedings of The IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 1055–1059.
[17] N. Ibtehaz, M.S. Rahman, MultiResUNet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation, Neural Networks 121 (2020) 74–87.
[18] Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. arXiv preprint arXiv:1809.10486.
[19] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial Transformer Networks, Advances in Neural Information Processing Systems 28 (2015) 2017–2025.
[20] Q. Jin, Z. Meng, T.D. Pham, Q. Chen, L. Wei, R. Su, DUNet: A Deformable Network for Retinal Vessel Segmentation, Knowledge-Based Systems 178 (2019) 149–162.
[21] X. Li, H. Chen, X. Qi, Q. Dou, C.W. Fu, P.A. Heng, H-denseunet: Hybrid Densely Connected UNet for Liver And Tumor Segmentation From CT Volumes, IEEE Transactions on Medical Imaging 37 (2018) 2663–2674.
[22] X. Li, W. Wang, X. Hu, J. Yang, Selective Kernel Networks, in: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.
[23] Y. Liu, D. Zhang, Q. Zhang, J. Han, Part-object relational visual saliency, IEEE Transactions on Pattern Analysis and Machine Intelligence Early Access (2021) 1.
[24] Z. Liu, Y.Q. Song, V.S. Sheng, L. Wang, R. Jiang, X. Zhang, D. Yuan, Liver CT Sequence Segmentation Based With Improved U-Net And Graph Cut, Expert Systems With Applications 126 (2019) 54–63.
[25] 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.
[26] F. Milletari, N. Navab, S.A. Ahmadi, V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, in: Proceedings of The IEEE Conference on 3D Vision, 2016, pp. 565–571.
[27] Z.L. Ni, G.B. Bian, X.H. Zhou, Z.G. Hou, X.L. Xie, C. Wang, Y.J. Zhou, R.Q. Li, Z. Li, Raunet: Residual Attention U-Net for Semantic Segmentation of Cataract Surgical Instruments, in: Proceedings of The International Conference on Neural Information Processing, 2019, pp. 139–149.
[28] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention U-Net: Learning Where to Look for the Pancreas. arXiv preprint arXiv:1804.03999.
[29] 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.
[30] Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A Large Annotated Medical Image Dataset for the Development and Evaluation of Segmentation Algorithms. arXiv preprint arXiv:1902.09063.
[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going Deeper With Convolutions, in: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[32] 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 Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[33] B. Wang, S. Qiu, H. He, Dual Encoding U-Net for Retinal Vessel Segmentation, in: Proceedings of The International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 84–92.
[34] C. Wang, Z. Zhao, Q. Ren, Y. Xu, Y. Yu, Dense U-Net Based on Patch-Based Learning for Retinal Vessel Segmentation, Entropy 21 (2019) 168.
[35] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual Attention Network for Image Classification, in: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.
[36] 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.
[37] X. Wang, R. Girshick, A. Gupta, K. He, Non-Local Neural Networks, in: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[38] S. Woo, J. Park, J.Y. Lee, I. So Kweon, CBAM: Convolutional Block Attention Module, in: Proceedings of The European Conference on Computer Vision, 2018, pp. 3–19.
[39] J. Xie, Z. Ma, D. Chang, G. Zhang, J. Guo, GPCA: A probabilistic framework for gaussian process embedded channel attention, IEEE Transactions on Pattern Analysis and Machine Intelligence Early Access (2021) 1.
[40] J. Xie, Z. Ma, J. Lei, G. Zhang, J.H. Xue, Z.H. Tan, J. Guo, Advanced dropout: A model-free methodology for bayesian dropout optimization, IEEE Transactions on Pattern Analysis and Machine Intelligence Early Access (2021) 1.
[41] Yu, F., Koltun, V., 2015. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv preprint arXiv:1511.07122.
[42] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Wang, Y. Yu, Exploring task structure for brain tumor segmentation from multi-modality mr images, IEEE Transactions on Image Processing 29 (2020) 9032–9043.
[43] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Yu, Cross-modality deep feature learning for brain tumor segmentation, Pattern Recognition 110 (2021) 107562.
[44] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-Attention Generative Adversarial Networks, in: International Conference on Machine Learning, 2019, pp. 7354–7363.
[45] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid Scene Parsing Network, in: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
[46] 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, Medical Physics 44 (2017) 5221–5233.
[47] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: A Nested U-Net Architecture for Medical Image Segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018, pp. 3–11.

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.

Shijie Liu 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 Biomedical Engineering from Hebei University of Technology, China, in 2019. His research interests lie in medical image processing using machine learning and deep learning methods.

Di Yuan is currently a PhD 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.

Lei Wang received the Ph.D. degree in theory and new technology of electrical engineering from the Hebei University of Technology, Tianjin, China, in 2009. He is currently a M.S. Superviser and an associate Professor with the Hebei University of Technology. His current research interests include neural engineering and brain science.

Junyang Chen received a Ph.D. degree in computer and information science from University of Macau, Macau, China, in 2020. He is currently an assistant professor with the College of Computer Science and Software Engineering, Shenzhen University, China. His research interests include graph neural networks, text mining, and recommender systems.

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.

Zhigang Fu is currently a Chief Physician in Chinese People's Liberation Army 983 Hospital. His current research interests mainly focus on telemedicine, physical examination, and health management.

Rui Zhang's research interests include AI and big data, particularly in the areas of recommendation systems, knowledge bases, chatbot, and spatial and temporal data analytics. He is a Visiting Professor at Tsinghua University and has previously been a Professor at the School of Computing and Information Systems of the University of Melbourne. Dr. Zhang has won the Future Fellowship by the Australian Research Council in 2012, Chris Wallace Award for Outstanding Research by the Computing Research and Education Association of Australasia (CORE) in 2015, and Google Faculty Research Award in 2017. Dr. Zhang obtained his Bachelor's degree from Tsinghua University in 2001 and PhD from National University of Singapore in 2006.