Constrained multi-scale dense connections for biomedical image segmentation

Jiawei Zhang [a,b], Yanchun Zhang [a,c,e], Hailong Qiu [b], Tianchen Wang [d], Xiaomeng Li [j], Shanfeng Zhu [f,g,h,i], Meiping Huang [b], Jian Zhuang [b], Yiyu Shi [d], Xiaowei Xu [b,*]

[a] Department of New Network, Pengcheng Laboratory, Shenzhen, Guangdong, China
[b] Guangdong Cardiovascular Institute, Guangdong Provincial Key Laboratory of South China Structural Heart Disease, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, 510080, China
[c] School of Computer Science, Zhejiang Normal Univeristy, Jinhua, China
[d] Department of Computer Science and Engineering, University of Notre Dame, IN, United States
[e] College of Engineering and Science, Victoria University, Melbourne, Australia
[f] Institute of Science and Technology for Brain-Inspired Intelligence and MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China
[g] Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, Shanghai, China
[h] Shanghai Key Lab of Intelligent Information Processing and Shanghai Institute of Artificial Intelligence Algorithm, Shanghai, China
[i] Zhangjiang Fudan International Innovation Center, Shanghai, China
[j] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

ARTICLE INFO

ABSTRACT

Multi-scale dense connection has been widely used in the biomedical image community to enhance the segmentation performance. In this way, features from all or most scales are aggregated or iteratively fused. However, by analyzing the details, we discover that some connections involving distant scales may not contribute to, or even harm, the performance, while they always introduce a noticeable increase in computational cost. In this paper, we propose constrained multi-scale dense connections (CMDC) for biomedical image segmentation. In contrast to current general lightweight approaches, we first introduce two methods, a naive method and a network architecture search (NAS)-based method, to remove redundant connections and verify the optimal connection configuration, thereby improving overall efficiency and accuracy. The results demonstrate that the two approaches obtain a similar optimal configuration in which most features at the adjacent scales are connected. Then, we applied the optimal configuration to various backbone networks to build constrained multi-scale dense networks (CMD-Net). Experimental results evaluated on eight image segmentation datasets covering biomedical images and natural images demonstrate the effectiveness of CMD-Net across a variety of backbone networks (FCN, U-Net, DeepLabV3, SegNet, FCNsa, ConvUNeXt) with a much lower increase in computational cost. Furthermore, CMD-Net achieves state-of-the-art performance on four publicly available datasets. We believe that the CMDC method can offer valuable insight for ways to engage in dense connectivity at multiple scales within communities. The source code has been made available at https://github.com/JerRuy/CMD-Net.

1. Introduction

Accurate segmentation of biomedical tissues, such as organs [1,2], tumors [3–5], or other anatomical structures [6–8], is a critical prerequisite for obtaining valid morphological statistics, which are extensively utilized in quantitative diagnosis and clinical decision-making. These morphological measurements, including size [3–5], shape [9–11], and texture features [6,7], provide valuable insights into the underlying pathological conditions and are essential for accurate diagnosis, prognosis, and treatment planning. As an example, Wang et al. [12] developed a deep learning-based system that could accurately detect lymph nodes and segment the tumor region in medical images of gastric cancer patients. By leveraging these automated segmentation capabilities, the system was able to provide valuable insights into the tumor characteristics and stage, which enabled more reliable prediction of the clinical outcome and prognosis for gastric cancer patients. This demonstrates
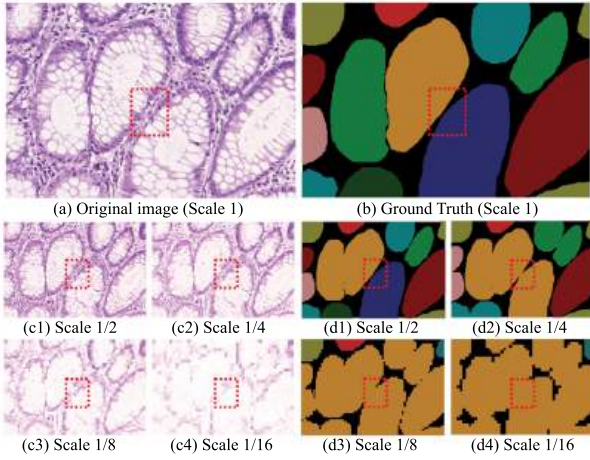
Fig. 1. Motivation explanation. (a) and (b) are an original gland tissue image and its mask, respectively, while (c1–c4) and (d1–d4) represent their scaled results. The masked regions with different colors correspond to different gland instances. (divided by path-connected domain) It can be noticed that most appearance information can be preserved in small-scale transformation, but large-scale transformation leads to this crucial information loss. As a result, proper or constrained connections that merely fuse context information from surrounding scales may obtain better segmentation results. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the potential of advanced deep learning techniques in enhancing the efficiency and accuracy of quantitative diagnosis and treatment planning for complex oncological conditions. However, manual segmentation of biomedical images and tissues carried out by professional clinicians is a time-consuming, costly, and inherently subjective process. The accuracy and consistency of such manual segmentation can be heavily influenced by factors such as the clinician's experience, fatigue, and individual interpretation. This introduces variability and potential errors, hindering the standardization and reliability of quantitative diagnostic measures derived from the segmented data. Consequently, there is a significant demand for automatic segmentation in clinical practice to enhance efficiency, reliability, and alleviate clinicians' burden.

Recently, fully convolutional networks (FCN) have been widely employed for automatic biomedical image segmentation [13]. Inspired by Dense-Net [14], current methods always employed multi-scale dense connection [15] to further enhance the segmentation performance. Different from Dense-Net, which fuses features at the same scale, multi-scale dense connections fuses features at multiple scales. In particular, features at almost all scales are fused to achieve optimal performance. However, this may be not the case. The connections at distant scales often introduce a large-scale feature transformation for spatial dimension alignment of features. For example, as shown in Fig. 1, it is difficult to separate tiny gaps between neighboring glands and adhesive glands from complex backgrounds, whose size is even smaller than the down-sampling size (marked by red dashed boxes). In addition, the large-scale downsampling process loses a lot of the appearance information, e.g., some neighboring glands are regard as the same gland (divided by path-connected domain). Furthermore, the large-scale up-sampling process usually yields non-smooth segmentation results and, in some cases, results in segmentation inconsistencies. Based on these observations, we have found that some connections with distant scales in dense networks may not contribute to or even harm the overall performance. Therefore, intuitively, establishing appropriate connections to fuse feature maps at multiple scales could potentially enhance overall segmentation performance.

Meanwhile, although it can be demonstrated that the improvement in model performance is not solely attributed to the increase in computational cost, the substantial rise in computational expenses

remains a concern. Concurrently, a variety of subsequent approaches employed a lot of lightweight approaches, including lightweight convolution [16], model distillation [17], and model pruning [18], model quantization [19], to mitigate the substantial increase in computational costs. However, these works actually start from general lightweight methods, and lack the essential thinking about the design of multi-scale dense connection.

Motivated by the preceding discussion, we introduce constrained multi-scale dense connections (CMDC) for accurate biomedical image segmentation. In CMDC, features are fused solely with adjacent scales containing pertinent appearance or semantic data. To determine the optimal connection configuration in CMDC, we introduce dense connection range, and employ a naive approach and a network architecture search (NAS) based approach for analysis and discussion. Experimental results show that both two approaches obtain a similar optimal connection configuration in which only the features at the adjacent scales are connected. We further applied CMDC to different locations (the encoder, the decoder, and their cross) of the segmentation networks, and propose the constrained dense networks (CMD-Net) for accurate biomedical image segmentation. The experimental results on four medical image segmentation datasets [3,6,20,21] demonstrate that CMD-Net achieves state-of-the-art segmentation performance among existing methods. Furthermore, we also executed extensive experiments for a variety of segmentation networks on eight image segmentation datasets [3,6,9,20–24] covering biomedical images and natural images to analyze the generalization of the CMDC. The experimental results demonstrate that our CMDC methods obviously increase segmentation performance across a variety of segmentation networks and datasets. Also, our method has a much lower computational cost, thus with high efficiency. In summary, our contributions are as follows:

- We observe that current multi-scale dense connections are not optimal. The connections at distant scales often introduce a large-scale feature transformation for spatial dimension alignment of features. Those connections may not contribute to or even harm the overall performance, which may lose appearance information and lead to segmentation inconsistency in some cases;
- To find the optimal connection configuration of dense connections, we introduce dense connection range, and adopt a naive approach and a network architecture search (NAS) based approach for discussion. Experimental results show that both the two approaches obtain a similar optimal connection configuration in which most of the features from the adjacent scales are connected, while most features at distant scales are discarded;
- Based on the experiments of the optimal connection configuration, we propose constrained multi-scale dense connections (CMDC) by merely aggregating feature maps at the adjacent scales, including relevant features of the current scale. Based on CMDC, we further propose the constrained dense networks (CMD-Net) by applying it to the encoder, the decoder, and their cross;
- Experiments on four biomedical image datasets shows that our CMD-Net achieves state-of-the-art performance against existing works. Furthermore, extended experiments on eight datasets covering biomedical image and natural image indicate that our CMDC improve segmentation performance effectively across various architectures and domains.

The remainder of the paper is organized as follows. Section 2 provides a brief overview of related studies. Section 3 goes into the proposed approach in depth, including CMDC, CMD-Net, and the dense connection range. Section 4 presents the experimental setup, method discussion, and result analysis. Section 5 presents the conclusion.
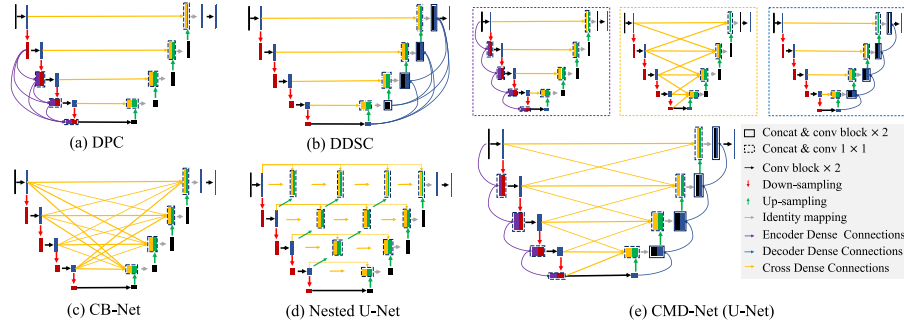
Fig. 2. Network structure comparison of (a-d) existing multi-scale dense connections methods (Dense Pooling Connections (DPC) [25], Dense Decoder Short Connections (DDSC) [26], Complete Bipartite Network (CB-Net) [27] and U-Net++ [28]) and (e) our proposed CMDC. (a-d) are representative methods including Dense Pooling Connections (DPC) [25], Dense Decoder Short Connections (DDSC) [26], Complete Bipartite Network (CB-Net) [27] and U-Net++ [28]. Purple, yellow and blue dotted line boxes are the *CMDC-based encoder*, the *CMDC-based decoder*, and the *CMDC-based cross*, which correspond to the architectures implementing CMDC to the encoder, the decoder, and their cross, respectively. In addition, CMD-Net (e) is the fusion of above three CMDC connections. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Related work

2.1. Multi-scale dense connection

As shown in Fig. 2(a–d), many recent researches have used multi-scale dense connections to relieve the vanishing gradient problem, enhance feature propagation, and increase feature reuse. There are mainly two approaches: direct fusion and iterative fusion. The former directly connects previous layers with the current layer using downsampling. For example, dense pooling connections [25] and dense decoder short connections [26] integrated full or most scales feature maps from previous layers on the encoder and the decoder, respectively. Inspired by bipartite graph, complete bipartite network [27] aggregated feature maps from the encoder to improve the segmentation performance. The later gradually adjusts the size of feature maps in previous layers and fuses them with the corresponding feature maps to refine the information fusion, which usually adds a large number of intermediate convolutions. U-Net++ [28] is a representative method of iterative fusion, which aggregated features at surrounding scales by connecting the encoder and the decoder through a series of nested and dense skip connections. Iterative fusion usually adds a large number of intermediate convolutions for performance improvement. Note that efficient dense connection [29] and hyper dense connection [30] perform feature fusion at the same scale (similar to Dense-Net), and thus are out of the scope of this paper. Fig. 2(e) shows the network structure of our CMD-Net, which has less connections in the encoder, the decoder and their cross compared with existing works.

As mentioned above, it can be concluded that multi-scale dense connections are effective on a wide range of datasets and basic models, but they introduce a greater increase in computing costs. Although it can be demonstrated that the improvement in model performance is not solely attributed to the increase in computational cost, the substantial rise in computational expenses remains a concern. Consequently, a variety of subsequent approaches employed many lightweight approaches, including lightweight convolution [16], model distillation [17], and model pruning [18], model quantization [19], to mitigate the substantial increase in computational costs. However, these works actually start from general methods of reducing the amount of computational cost, and lack the essential thinking about the design of multi-scale dense connection.

Our initial conference paper [31] introduced an intuitive naive design principle to find the optimal configuration of dense connections, which is concise and effective. However, this approach is subjective and hand-crafted, and the experiments were limited to medical images. Furthermore, this manuscript has added an objective computer simulation verification based on network architecture search to verify the effectiveness of CMDC. We believe that this can offer valuable insight into ways to engage in dense connectivity at multiple scales within communities. Furthermore, we have added four other widely used datasets, including two biomedical image segmentation datasets (DRIVE [6] and CHASE-DB1 [21]) and two natural image segmentation datasets (VOC2012 [23] and Cityscapes [24]) in the experiment to further verify the effectiveness of the CMDC.

2.2. Network architecture search

In recent years, neural architecture search (NAS) has demonstrated great achievements in designing neural network architectures automatically and achieved mass performance improvements in various tasks. Meanwhile, NAS with multi-scale feature aggregation has become a popular paradigm. For example, Dense-NAS [32] designed a densely connected search space to build densely connected routing blocks. There are also some works particularly for biomedical images. Most of them are based on the classic network U-Net [33], which is the most widely-used network for biomedical image segmentation. For example, NAS U-Net [34] searched the cell architecture in the U-shape architecture for biomedical tasks, while MS-NAS [35] employed a larger search space at different scales (including network level and cell level) to improve segmentation performance. In summary, the above-mentioned NAS methods for medical images are mostly based on basic unit structure search of classic structures (such as U-Net, etc.) or search of the overall architecture. The search in our work is dedicated to whether connections between different scale features in existing classical structures are beneficial to the learning of the overall network. This is obviously different from existing NAS methods and is also a positive supplement.

2.3. Biomedical image segmentation

The field of medical imaging has witnessed remarkable progress in recent years, largely driven by the advancements in deep learning techniques. Currently, fully convolutional networks (FCNs) including the U-Net architecture [33] have dominated biomedical image segmentation, finding applications in tasks such as gland image segmentation [3,20], esophageal cancer image segmentation [22], wireless capsule endoscopy (WCE) image bleeding area segmentation [9], liver CT segmentation [2], and vessel image segmentation [7,21]. For example, the DCAN model [4] created a contour recognition decoding branch for unified multi-task learning, with well-defined object boundaries. Beyond the U-Net architecture, researchers have proposed various improvements and novel network designs for medical image segmentation. Han et al. [36] introduced the ConvUNeXt model, an efficient convolutional neural network, while Jafari et al. [37] presented the DRU-Net, an efficient deep convolutional neural network. Further

advancements include the Dense-PSP-UNet by Ansari et al. [38], which combines dense connections and pyramid pooling for fast and accurate liver ultrasound segmentation. The integration of transformers with convolutional neural networks has also shown promising results in 3D medical image segmentation, as demonstrated by the Cotr model proposed by Xie et al. [39]. Active learning approaches have been intensively investigated to reduce the work of human annotations, with methods like suggestive annotation [40] actively picking the most representative examples based on uncertainty and similarity. The MILD-Net [20] introduced a complex structure with a minimal information loss unit to compensate for data loss during down-sampling. Additionally, the work in [9] designed a multi-stage architecture and attention blocks to deal with small area segmentation in WCE images. Regarding vessel segmentation, more methods have focused on incorporating multi-scale context information to improve the accuracy of thick and thin vessel segmentation, such as the pyramid scale aggregation block [41] and the separation of thick and thin vessels in different branches [42]. Beyond image segmentation, deep learning techniques have also been employed for risk assessment and practical utility in clinical settings, as explored by Akhtar et al. [43] and Ansari et al. [44,45]. In this paper, we compare the aforementioned state-of-the-art biomedical image segmentation methods across various medical domains to demonstrate the superior performance of the proposed CMD-Net.

3. Methods

We introduce CMDC in this part by applying it to the encoder, the decoder, and their cross, and then fuse them to build CMD-Net with demonstrations on a variety of existing networks. For ease of explanation, we use U-Net as a vehicle in the following discussion.

For ease of discussion of the connection configuration in CMDC, we define dense connection range as the number of surrounding scales that needs to be fused by the current scale. For example, setting it to 2 means each scale needs to consider five scales (two surrounding higher scales, one current scale, and two surrounding lower scales). **For the convenience of presentation, the dense connection range is set to 1 in this section.** Note that the definition of dense connection range has the following two intuitions/constraints introduced by our motivation. First, the connection is symmetrical, e.g., if features at the adjacent higher scale is connected, features at the adjacent lower scale (if exists) should be connected also. Second, features at surrounding scales are more important to the current scale than that at distant scales, e.g., setting the dense connection range to 2, then features at adjacent are also connected. We adopted a naive approach and a network architecture search (NAS) based approach to find the optimal dense connection range, which is detailed in Section 3.3. Note that in the NAS-based approach, the above two intuitions/constraints are not applied so that it can find the optimal configuration with high flexibility.

3.1. Constrained multi-scale dense connections

CMDC based Encoder/Decoder:
The construction of a traditional encoder is illustrated within the dashed rectangle in Fig. 3, where X_{i-1} and X_i denote the current layer's input and output feature maps, respectively. To enhance the input feature map using X_{i-2} from an adjacent scale, we replace X_d^{i-1} with X_{newd}^{i-1}, formed by concatenating X_d^{i-1} and X_{new}^{i-2}. Here, X_{new}^{i-2} represents down-sampled feature maps of X^{i-2}, and $H(\cdot)$ denotes feature concatenation followed by 1×1 convolution, which aggregates channel-wise context information while maintaining consistent channel counts across layers.
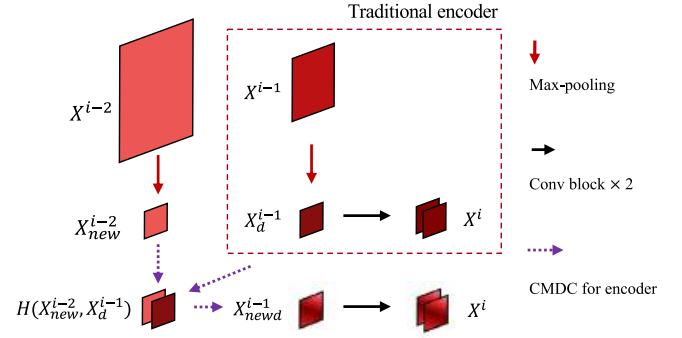
$$X_{newd}^{i-1} = H(X_{new}^{i-2}, X_d^{i-1}). \tag{1}$$

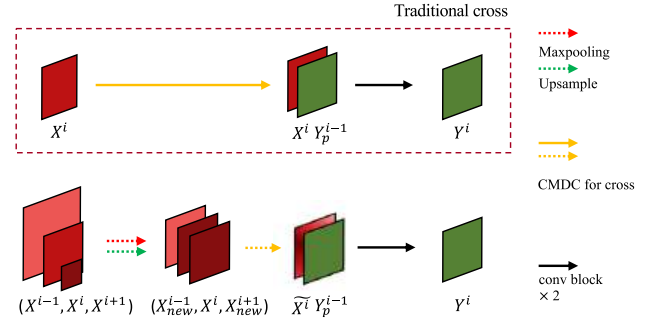**Fig. 3.** Comparison of our CMDC-based encoder and the traditional encoder.

Fig. 4. Illustration of our CMDC-based cross and the traditional cross depicting in the dashed boxed.

Similarly to the CMDC-based encoder, the CMDC-based decoder is defined as follows:

$$Y_{newp}^{i-1} = H(Y_{new}^{i-2}, Y_p^{i-1}), \tag{2}$$

Here, Y_{newp}^{i-1} represents enhanced results of Y_p^{i-1} achieved by aggregating feature maps from adjacent scales within the decoder.

CMDC based Cross of Encoder and Decoder: Fig. 4 illustrates the CMDC-based cross of the encoder and the decoder. Y^{i-1} and Y^i represent the current layer's input and output, respectively. Y_p^{i-1} is the intermediate feature maps of Y^{i-1} after up-sampling. In practice, U-Net encoders merely fuse feature maps from the same scale X^i to enhance features in the decoder, but CMDC encoders and decoders fuse feature maps from not only the same scale but also the adjacent scales. The CMDC-based cross fuses feature maps from three scales: the identical scale, the adjacent higher scale and the adjacent lower scale. $\widetilde{X}^{i-1}$ is employed to replace X^{i-1}, and $\widetilde{X}^{i-1}$ is calculated as,

$$\widetilde{X}^{i-1} = H(X_{new}^{i-1}, X^i, X_{new}^{i+1}). \tag{3}$$

Overall, CMDC merely replaces the original feature at each scale with the enhanced features, leaving the original convolution layer with the same number of channels.

Fig. 4 depicts the CMDC-based interaction between the encoder and the decoder. Here, Y^{i-1} and Y^i denote the input and output feature maps of the current layer, respectively, while Y_p^{i-1} represents the intermediate feature maps of Y^{i-1} post up-sampling. In contrast to U-Net encoders, which typically merge feature maps exclusively from the same scale X^i to enhance decoder features, CMDC encoders and decoders fuse feature maps not only from the same scale but also from adjacent scales. Specifically, the CMDC-based interaction integrates feature maps across three scales: the current scale, the adjacent higher scale, and the adjacent lower scale. $\widetilde{X}^{i-1}$ replaces X^{i-1} and is defined as follows:

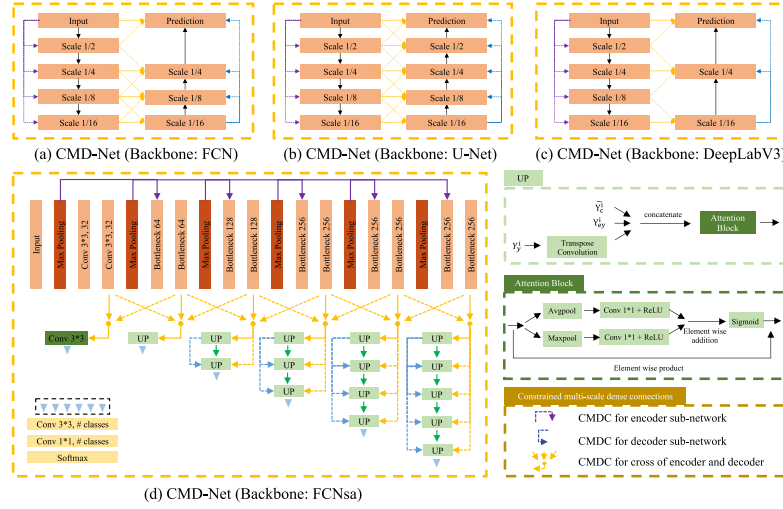$$\widetilde{X}^{i-1} = H(X_{new}^{i-1}, X^i, X_{new}^{i+1}). \tag{3}$$

Fig. 5. Network structure of CMD-Net family. A series of CMD-Net are generated by combining CMDC with various segmentation network backbones. For easy of discussion, the dense connection range is set to 1.

In summary, CMDC replaces original features at each scale with enhanced features, maintaining the convolutional layer's original channel count.

3.2. Constrained multi-scale dense networks

CMD-Net integrates components from the CMDC-based encoder, CMDC-based decoder, and their interaction cross. The details of each component's implementation have been extensively covered in previous sections, particularly focusing on the fusion between the CMDC-based encoder and the cross. It is important to note that, for simplicity, the dense connection range is set to 1. To clarify, we denote the output feature maps of the CMDC-based cross and decoder as Y_{newp}^{i-1} and $\widetilde{X}^{i-1}$, respectively. Y_p^{i-1} represents the output features of the transpose operation within the decoder, which are replaced by the enhanced feature $\Psi(\widetilde{\mathbf{Y}}^{i-1})$. $\widetilde{\mathbf{Y}}^{i-1}$ combines Y_p^{i-1}, Y_{newp}^{i-1}, and $\widetilde{X}^{i-1}$. Here, $\Psi(\cdot)$ signifies the channel attention operation, enhancing feature concatenation through inter-channel relationships. The channel information is fused using the average-pooling function $\Phi_{\mathbf{Avg}}(\cdot)$ and the max-pooling function $\Phi_{\mathbf{Max}}(\cdot)$. $\mathbf{W}(\cdot)$ denotes a 1×1 convolution with ReLU activation, while $\otimes$ denotes the Hadamard product. $\sigma(\cdot)$ represents the Sigmoid function. The channel attention operation $\Psi(\cdot)$ can be formulated as follows in its entirety:

$$\Psi(\widetilde{\mathbf{Y}}^{i-1}) = \sigma(\mathbf{W}(\Phi_{\mathbf{Avg}}(\widetilde{\mathbf{Y}}^{i-1})) + \mathbf{W}(\Phi_{\mathbf{Max}}(\widetilde{\mathbf{Y}}^{i-1}))) \otimes \widetilde{\mathbf{Y}}^{i-1}, \tag{4}$$

As demonstrated in Fig. 5, with modest adjustments, CMD-Net can be simply implemented based on existing backbones such as FCN-8s [13], U-Net [33], and DeepLab-V3 [46].

3.3. Dense connection range

We introduce two methods to search the optimal setting of the dense connection range, a naive method and a NAS-based method. The naive method directly evaluates the network performance with various dense connection ranges. Meanwhile, the NAS-based method trains a supernet that connects all possible scales and evaluates the contribution of each connection. Thus, it is an empirical, time-consuming and un-efficient process but with some sort of generalization and interpretability. It is an automatic process thus with high efficiency. However, the obtained models usually have irregular connections, thus with low generalization and interpretability. We employ both methods to find the optimal
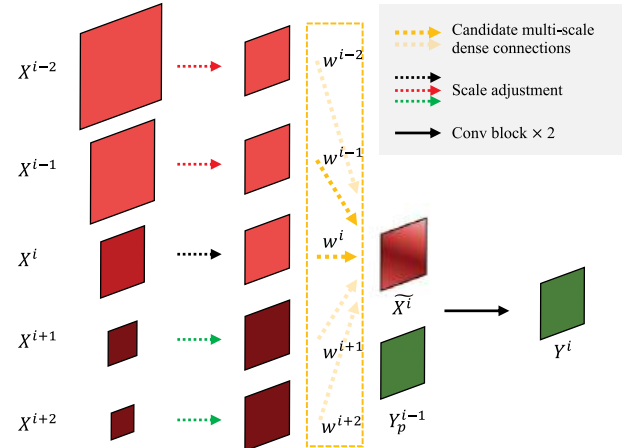

Fig. 6. Example of the search space of dense connection range in the cross. The yellow rectangle represents candidate multi-scale connections. The hyper-parameter w^i is defined as the importance of the connection to the ith scale in the NAS process. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

configuration of dense connection range. Fig. 6 illustrates an example search space of dense connection range in the cross. The network has a depth of L, which means the encoder has L scales. In the ith scale of the decoder, feature maps in each scale are scaled into the ith scale for feature concatenation. The hyper-parameter w^i presents the contribution of X^i from the i scale in the encoder, which are normalized as $Softmax(w^i) = \frac{\exp\{w^i\}}{\sum_{i \in L} \exp\{w^i\}}$. The overall process can be described as,

$$\widetilde{x^i} = H([w^i \cdot X^i]_{i \in L}), \tag{5}$$

where $H(\cdot)$ denotes feature concatenation. Note that the first layer in the encoder and the last layer in the decoder can only aggregate one connection, and thus we do not implement the softmax function in these layers. The output $\widetilde{x^i}$ of the multi-scale dense connections is then computed as the concatenation of the scaled feature maps from different scales in the encoder.

Table 1
Information of biomedical image datasets used in our experiments.

Datasets	Lesion	Train	Test	Resolution	Modality
GlaS [3]	Colon	85	80	775 × 522	Histopathology slides
CRAG [20]	Colon	173	40	1512 × 1516	Histopathology slides
BISW [9]	Bowel	235	100	360 × 360	Endoscopy
ESOS [22]	Esophagus	466	64	830 × 1436	Histopathology slides
DRIVE [6]	Fundus	20	20	565 × 584	Fundus photography
CHASE-DB1 [21]	Fundus	20	8	999 × 960	Fundus photography

Table 2
Information of natural image datasets used in our experiments.

Datasets	Categories	Train	Validation	Test
VOC2012 [23]	21	10 582	1449	1456
Cityscapes [24]	19	2975	500	1525

4. Experiments

4.1. Experimental setup

Datasets. In the experiment, we used eight image segmentation datasets covering medical images and natural images to verify the effectiveness of CMDC and CMD-Net. In terms of biomedical image, we use six biomedical image datasets for evaluation covering various medical imaging modalities, which is detailed in Table 1. In terms of natural image, we use two widely-used natural image datasets for evaluation, which are detailed in Table 2. We also proposed a more comprehensive analysis demonstrating the dataset's relevance and applicability, which can be find in supplementary material.

Implementations. Our proposed solutions were tested on two NVIDIA GeForce GTX TITAN X (pascal) graphics cards, each with 12 GB of RAM. In the beginning of the training, we used a batch size of four and a learning rate of 0.005. Besides, we chose the Adam optimizer and cross-entropy loss to optimize the network. Furthermore, we utilized the same configuration to ensure fair comparisons. Considering the fact that the gland tissue is continuous and smooth, the raw segmentation from CMD-Net was refined through a disk filter in post-processing to smooth the segmented tissue mask, fill gaps, and eliminate tiny areas.

For NAS, we implemented the search on the network with a depth of 4. Similar to [47], we defined two sets of parameter, α and β, to optimize to obtain an approximate solution. α includes the hyper-parameters of the multi-scale dense connection weighting each connection contribution in each layer, while β contains other parameters. The two sets are optimized by the Adam optimizer, and the learning rates for α and β are 0.005 and 0.001, respectively. The search space of the encoder has $2^1 + 2^2 + 2^3 + 2^4 = 31$ unique paths, while the search space of the decoder is the same as that of the encoder. Meanwhile, the search of the cross has $2^5 + 2^5 + 2^5 + 2^5 = 128$ unique paths. Thus, the search space has $31 \times 128 \times 31 = 123\,008$ unique paths in total.

4.2. Method discussion

In this section, we focus on the optimal setting of multi-scale dense connections, including dense connection range and CMDC location.

4.2.1. Dense connection range

As discussed in Section 3.3, we introduce two methods to configure the optimal setting of the dense connection range, a naive method and a NAS-based method, which are both implemented based on U-Net.

Naive CMD-Net: In Fig. 7, the ablation results on the GlaS dataset from the original U-Net and one with various dense connection range on the encoder, the decoder, and the cross are shown. Overall, CMDC can significantly increase instance segmentation performance, often by more than 10% on Object Dice score. Meanwhile, when the dense connection range grows, the accuracy swings in a limited range, without
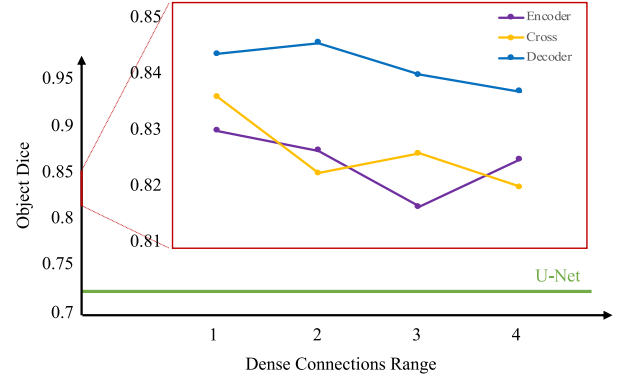

Fig. 7. Ablation experiment of the dense connection range and locations of CMDC on the GlaS dataset test A part.

a significant gain or even a minor drop. This illustrates that fusion of too many scales does not improve the network performance, which is consistent with our intuition in Section 1.

Furthermore, we conducted more experiments in Naive-based CMDC to directly verify the relationship between dense connections and scales. As shown in Fig. 8, from left to right, we conducted detailed experimental analysis on the dense connections between the feature in the encoder, between the feature cross the encoder and the decoder, and between the feature in the decoder, and the different scales of the connection. As shown in Fig. 7a, ei (i belongs to [1,2,3,4]) represents that in the encoder of the dense connection set we designed, only the features with a difference of i scale transformations are connected for feature fusion. We can find that the dense connection that only connects the features with a difference of 1 scale transformation (actually represents only the features of the nearest scale) is significantly higher than the others. Meanwhile, similar experimental results were observed in the experiments on the cross between the encoder and decoder, and the decoder. This also shows that in the multi scale dense connection, the dense connection of the nearest scale feature has the highest benefit. Both experimental results demonstrate that dense connections can bring performance gains, but for efficiency reasons, cross-scale dense connections that only connect the nearest neighbors can bring the best efficiency and performance benefits. **Thus, we obtain the optimal configuration in our naive CMD-Net that the dense connection range is set to 1 in the encoder, the decoder and their cross.**

NAS-based CMD-Net: We implemented the NAS-based method on the GlaS dataset to search the optimal configuration of multi-scale dense connections. Each connection is weighted by a hyper-parameter to evaluate the importance of the connection. Fig. 9 illustrates the NAS-based CMD-Net on the GlaS dataset. It can be noticed that although the search results have a certain randomness, **most of the remained connections in the NAS-based CMD-Net are with adjacent scales (the dense connection range is 1)**. This also shows that the network does not benefit from the connection with distant scales.

Moreover, we conducted further experiments on NAS-based CMDC to independently verify the search results of dense connections in different network parts. As shown in Fig. 10, from left to right, we obtained
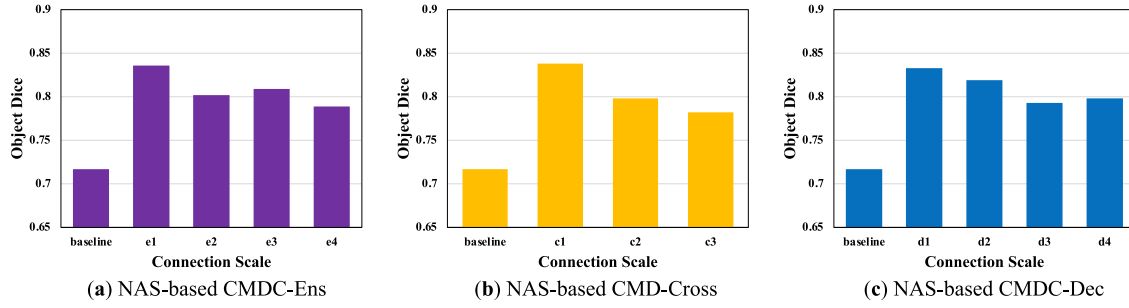
Fig. 8. Ablation experiment of the dense connection scale and locations of CMDC on the GlaS dataset test A part.
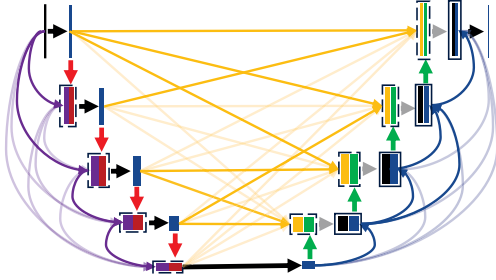

Fig. 9. Illustration of multi-scale dense connections searched on the GlaS dataset with a depth of 4. Best view in color.

Table 3

Comparison of Naive CMD-Net and NAS CMD-Net on (a) the GlaS dataset and (b) the CRAG dataset. The connection similarity is defined as the ratio of the number of the connections with adjacent scale (constrained) and the number of all connections.

Method	Connection similarity	Object Dice
Naive CMD-Net	100.0%	84.8%
NAS-based CMD-Net	76.5%	85.3%

Table 4

Ablation experiments of dense connection locations on the GlaS dataset and the CRAG dataset. The original U-Net, U-Net with CMDC under different configurations, and other multi-scale dense connections approaches are compared in terms of **instance** segmentation performance. Both accuracy and efficiency are compared.

Method	Object F1		Object Dice		Efficiency		
	GlaS	CRAG	GlaS	CRAG	Mem.	Params.	FLOPs
U-Net [33]	61.9%	60.0%	71.7%	65.4%	4.3	7.76	158.67
DPC [25]	78.1%	77.5%	82.2%	79.0%	4.9	9.08	166.20
CMDC-Enc	**80.2%**	**79.1%**	**83.6%**	**80.8%**	**4.6**	**7.90**	**160.70**
U-Net++ [28]	73.3%	74.1%	79.2%	75.3%	10.4	9.05	391.72
CB-Net [27]	76.0%	77.8%	81.8%	**78.5%**	–	**8.00**	198.75
CMDC-Cross	**79.1%**	**77.9%**	**83.8%**	78.1%	**7.3**	8.07	**178.02**
DDSC [26]	80.1%	77.6%	82.1%	76.9%	6.8	10.35	314.71
CMDC-Dec	**80.7%**	**78.6%**	**83.3%**	**77.7%**	**6.2**	**8.42**	**173.66**
CMDC U-Net	**81.9%**	**81.1%**	**84.8%**	**81.9%**	9.8	8.86	195.24

- means that cannot implement on a single GPU TITAN X (pascal) with 12 GB of RAM.
* E, C and D indicate that CMDC is applied to the encoder, the decoder and their cross, respectively.

the search results of dense connections in the encoder, between the encoder and the decoder, and the decoder. We can find that the frequency of dense connections that only connect features with a difference of 1 scale transformation (actually representing only connecting features of the nearest scale) in the search results is significantly higher than that of others. At the same time, similar experimental results were observed in the decoder and between the encoder and the decoder. This also shows that in the multi scale dense connection, the dense connection of the nearest scale feature has the highest benefit.

Overall Discussion: For analysis, we define connection similarity of the NAS-based CMD-Net and our naive CMD-Net as the ratio of the number of the connections with adjacent scales (constrained) and the number of all connections. Table 3 summarizes the connection similarity and accuracy of our naive CMD-Net and NAS-based CMD-Net. We can notice that NAS-based CMD-Net has a high connection similarity of 76.5%. More details of the connections in the NAS-based CMD-Net is shown in Fig. 9, in which 13 of 17 connections (76.5%) belong to adjacent scales. Note that the multi-scale densely connected networks with a depth of 4 have 36 connections, and 18 of 36 (50.0%) are adjacent scales. Compared with multi-scale densely connected networks, such a high connection similarity in the NAS-based CMD-Net shows a similar phenomenon in our naive CMD-Net. Thus, connections from adjacent scales are more important than that from distant scales, which is consistent with the phenomenon in our Naive CMD-Net.

Table 3 also shows that the naive CMD-Net achieves comparable results with the NAS-based CMD-Net on the GlaS dataset. Although the NAS-based CMD-Net achieves a slightly higher accuracy with a certain of randomness, it lacks generalization and interpretability. In

addition, it consumes much more training time to obtain a NAS-based CMD-Net. Therefore, in practice, we recommend to simply adopt a dense connection rage of 1 in the network configuration. **For ease of discussion, we define CMD-Net as the naive CMD-Net with dense connection rage of 1 in the rest of the paper.**

4.2.2. Dense connection locations

We explored CMD-Net with various dense connection locations and compared our approaches with existing works including DPC [25], CB-Net [27], U-Net++ [28], and DDSC [26]. Table 4 shows the ablation experimental results on the GlaS dataset and the CRAG dataset. Compared with existing works, our methods improve the accuracy by 1.7%, 1.2%, and 0.9% on average by using the CMDC-based encoder, the CMDC-based decoder, and their cross, respectively. Compared with merely using CMDC on the encoder, the decoder, and their cross, the combination of the three can provide much better results, with an average improvement of 2.15%. We can also discover that the network with the CMDC-based encoder had a higher average improvement than that with the CMDC-based decoders (0.85%) and their cross (1.2%). It might be due to the fact that early feature sharing and reuse might have a larger influence across the whole networks, resulting in better performance.

In addition, our method consumes much fewer resources than existing works. We can notice that the CMDC-based encoder/decoder/cross consume more resources than U-Net (on average 40.3%, 4.5%, and 7.6% more resources on memory, parameters, and FLOPs, respectively), and also achieve much higher performance than U-Net. Compared with existing networks, the CMDC-based encoder/decoder/cross achieve a decrease in memory, parameters, and FLOPs by 20%, 16.7%, and 70.0% respectively.
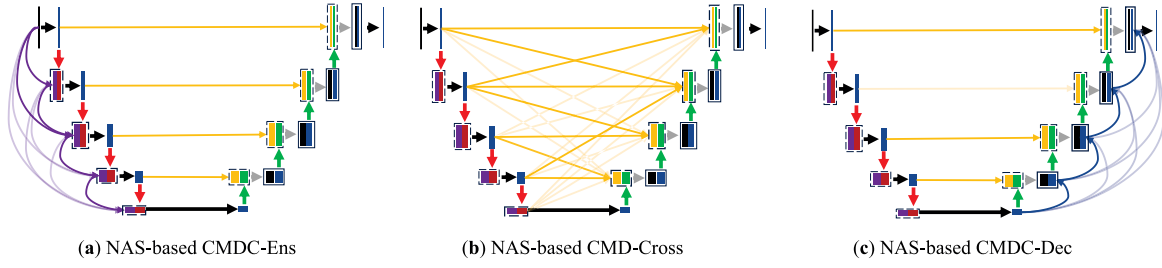
(a) NAS-based CMDC-Ens   **(b)** NAS-based CMD-Cross   **(c)** NAS-based CMDC-Dec

Fig. 10. Illustration of multi-scale dense connections searched on the GlaS dataset with a depth of 4, in the encoder, the decoder, and the cross of them, respectively. Best view in color.
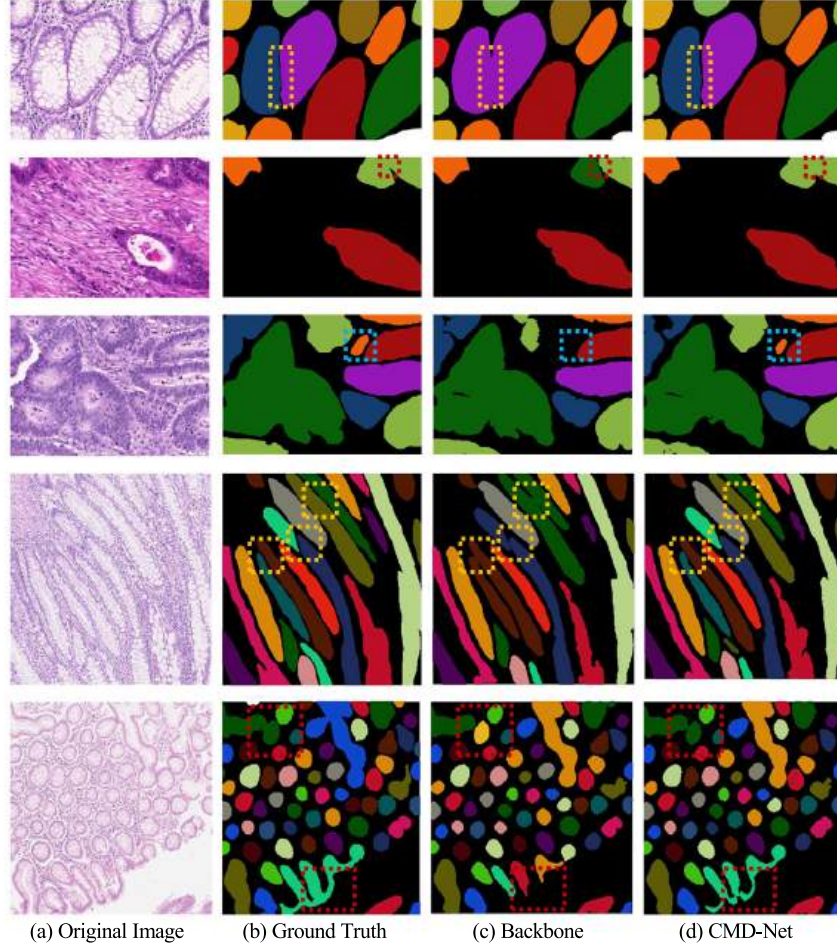

(a) Original Image   (b) Ground Truth   (c) Backbone   (d) CMD-Net

Fig. 11. Results visualization of CMD-Net and its backbone model on gland segmentation. Images in Row 1–3 and Row 4–5 are from the GlaS dataset and the CRAG dataset, respectively. Rectangle boxes represent some commonly-seen instance segmentation situations including false merging (yellow), false split (red), and false negative (blue). Note that erroneous merging has a big negative impact on the performance of gland instance segmentation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3. Results and discussion

With the optimal connection configuration (the dense connection range equals 1) and the optimal dense connection locations (applying CMDC to the encoder, the decoder and their cross) as discussed in the previous subsection, we further implemented CMD-Net and conducted extensive experiments across various backbones (including FCN-8s [13], U-Net [33], DeepLabV3 [46], and FCNsa [40]) and datasets (including GlaS [3], CRAG [20], BISW [9], ESOS [22], DRIVE [6], and CHASE-DB1 [21]). The results and discussion are as follows.

4.3.1. Comparison with state-of-the-art works

In this part, we introduce the comparison of CMD-Net with prior works on the GlaS dataset and the CRAG dataset using metrics including Object F1-score, Object Dice coefficient, and Object Hausdorff distance, respectively. We first introduce the result analysis on GlaS dataset, which is shown in Table 5. On GlaS Test A, GCSBA-Net achieves a competitive result on a Object F1-score of 91.6%, a Object Dice of 91.4%, and a Object Hausdorff of 40.13. Meanwhile, CMD-Net achieves a state-of-the-art result on Object F1-score (increased by 0.3%) and Object Hausdorff (reduced by 1.36), as well as a competitive result on Object Dice (only 0.2% poorer). While on GlaS Test B, CMD-Net outperforms GCSBA-Net by 2.8%, 1.6%, and 4.56 on Object F1-score,
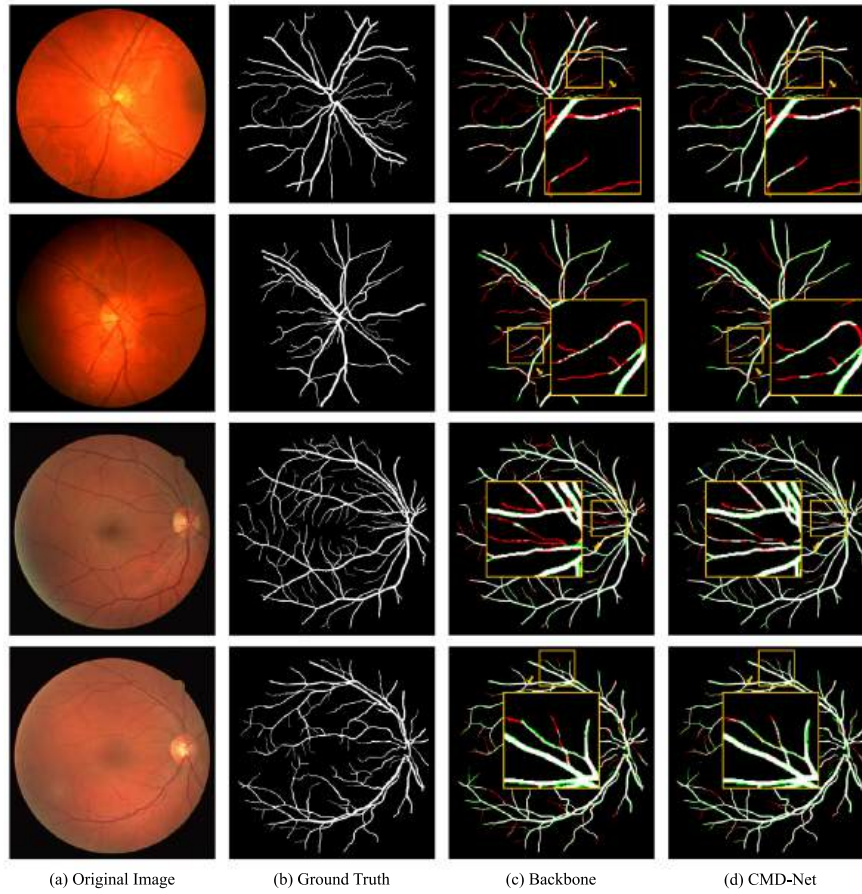
(a) Original Image   (b) Ground Truth   (c) Backbone   (d) CMD-Net

Fig. 12. Results visualization of CMD-Net and its backbone model on vessel segmentation. Images in Row 1–2 and Row 3–4 are from the DRIVE dataset and the CHASE-DB1 dataset, respectively.

Table 5
Comparison of CMD-Net with existing methods on the GlaS dataset.

Method	Object F1-score		Object Dice		Object Hausdorff	
	TestA	TestB	TestA	TestB	TestA	TestB
vision4GlaS	63.5%	52.7%	73.7%	61.0%	107.49	210.10
CUMedVision1	86.8%	76.9%	86.7%	80.0%	74.60	153.65
Freidburg2	87.0%	69.5%	87.6%	78.6%	57.09	148.47
ExB3	89.6%	71.9%	88.6%	76.5%	57.36	159.87
ExB1	89.1%	70.3%	88.2%	78.6%	57.41	145.58
CUMedVision2	91.2%	71.6%	89.7%	78.1%	45.42	160.35
FCN-8s [13]	78.3%	69.2%	79.5%	76.7%	105.04	147.28
SegNet [48]	85.8%	75.3%	86.4%	80.7%	62.62	118.51
DeepLab-v3 [46]	86.2%	76.4%	85.9%	80.4%	65.72	124.97
DCAN [4]	91.2%	71.6%	89.7%	78.1%	45.42	160.35
MILD-Net [20]	91.4%	84.4%	91.3%	83.6%	41.54	105.89
GCSBA-Net [5]	91.6%	83.2%	**91.4%**	83.4%	41.49	102.88
CMD-Net	**91.9%**	**86.0%**	91.2%	**84.8%**	**40.13**	**98.32**

Table 6
Performance comparison of CMD-Net with existing works on the CRAG dataset.

Method	Object F1-score	Object Dice	Object Hausdorff
FCN-8s [13]	55.8%	64.0%	436.43
SegNet [48]	62.2%	73.9%	247.84
DeepLab-v3 [46]	64.8%	74.5%	281.45
DCAN [4]	73.6%	79.4%	218.76
MILD-Net [20]	82.5%	87.5%	160.14
GCSBA-Net [5]	83.6%	**89.4%**	146.77
CMD-Net	**84.0%**	87.9%	**132.38**

Table 7
Performance comparison of CMD-Net with existing methods on the DRIVE dataset.

Method	Sensitivity	Specificity	Accuracy	AUC
FCN [13]	74.89%	96.21%	94.13%	95.67%
U-Net [33]	75.31%	96.45%	94.45%	96.01%
Three-stage [42]	76.31%	98.20%	95.38%	97.50%
BTS-DSN [49]	78.91%	98.04%	95.61%	98.06%
DEU-Net [50]	79.40%	9816%	95.67%	97.72%
Vessel-Net [8]	80.38%	98.02%	95.78%	98.21%
HA-Net [51]	79.91%	**98.13%**	95.81%	98.23%
CMD-Net	**84.56%**	96.92%	**95.96%**	**98.29%**

Table 8
Comparison of CMD-Net with existing methods on the CHASE-DB1 dataset.

Method	Sensitivity	Specificity	Accuracy	AUC
FCN [13]	76.41%	98.06%	96.07%	97.76%
BTS-DSN [49]	78.88%	98.01%	96.27%	98.40%
DEU-Net [50]	80.74%	98.21%	96.61%	98.12%
Vessel-Net [8]	81.32%	98.14%	96.61%	98.60%
Three-stage [42]	76.41%	98.06%	96.07%	97.76%
HA-Net [51]	**82.39%**	98.13%	96.70%	98.70%
CMD-Net	77.03%	**98.54%**	**96.93%**	**98.89%**

Object Dice, and Object Hausdorff, respectively. It should be noted that Test B is substantially more difficult than Test A due to the presence of more malignant patients. CMD-Net surpasses previous methods on the CRAG dataset by 0.4% on Object F1-score and 14.39 on Object Hausdorff. GCSBA-Net achieves good results using the Gabor-based module and the Cascade Squeeze Bi-Attention module on the CRAG

dataset, with an Object Dice of 89.4%, which is 1.5% higher than our CMD-Net.

Then, we introduce the result analysis on GlaS dataset, which is shown in Table 6. MILD-Net demonstrates significant advancement with an F1-score of 82.5% and a Dice coefficient of 87.5%, suggesting improved accuracy in object segmentation tasks compared to earlier methods. However, its Hausdorff distance of 160.14 indicates some variability in boundary prediction. Meanwhile, GCSBA-Net further enhances performance with an F1-score of 83.6% and a leading Dice coefficient of 89.4%, indicating superior overlap between predicted and ground truth objects. Its Hausdorff distance of 146.77 indicates a relatively precise boundary prediction. Next, CMD-Net emerges as the top performer in this comparison, achieving the highest F1-score of 84.0% among all methods. Although its Dice coefficient of 87.9% is slightly lower than GCSBA-Net, CMD-Net excels with the lowest Hausdorff distance of 132.38, indicating more accurate boundary localization compared to all other methods. CMD-Net demonstrates competitive performance across all metrics, showcasing robust segmentation accuracy and boundary delineation capabilities. While GCSBA-Net achieves a marginally higher Dice coefficient, CMD-Net's superior Hausdorff distance underscores its precise boundary prediction capability. Compared to MILD-Net, CMD-Net achieves a higher F1-score and exhibits more accurate boundary predictions, making it a preferred choice for biomedical image segmentation tasks on the CRAG dataset.

Visualization comparison of the segmentation results on gland segmentation is displayed in Fig. 11. In the boxed portions of Row 1 and Row 4, we can see that neighboring objects are extremely near to each other, and it is difficult to separate the tiny gaps between these neighboring glands. For example, the backbone approach cannot separate the two objects in Row 1 and accidentally segments a gland tissue into numerous small objects. In comparison to backbone model, CMD-Net is capable of handling these situations successfully, particularly at the object level. CMD-Net can also help with the segmentation of small glands (Row 3) and lumens of various forms (Row 2 and Row 5).

Meanwhile, two widely used vessel image segmentation datasets including DRIVE and CHASE-DB1 are also considered for comparison. The assessment results of CMD-Net and existing works are summarized in Tables 7 and 8. We can discern the performance characteristics of each method, particularly focusing on Sensitivity, Specificity, Accuracy, and AUC.

On DRIVE dataset, starting with sensitivity, CMD-Net achieves the highest value at 84.56%. This indicates CMD-Net's ability to effectively detect retinal vessels compared to other methods such as FCN (74.89%) and U-Net (75.31%), demonstrating its superior sensitivity in capturing true positive vessel pixels. Moving to specificity, HA-Net leads with 98.13%, closely followed by Vessel-Net at 98.02%. CMD-Net achieves a specificity of 96.92%, indicating its competency in minimizing false positives while identifying background pixels, although slightly lower than HA-Net and Vessel-Net. In terms of accuracy and AUC, CMD-Net achieves a accuracy of 95.96% and a AUC of 98.29%, which represents the overall correctness in classifying both vessel and non-vessel pixels. This places CMD-Net among the top performers, underscoring its robustness in providing accurate segmentation results. Those two metrics are crucial for evaluating the overall performance of the segmentation model. In conclusion, CMD-Net distinguishes itself with superior sensitivity and competitive overall performance on the DRIVE dataset, demonstrating its efficacy for precise retinal vessel segmentation tasks.

On the CHASE-DB1 dataset, starting with Sensitivity, which measures the ability to correctly identify vessel pixels, CMD-Net achieves 77.03%. This value is competitive, albeit slightly lower compared to HA-Net at 82.39%, which demonstrates HA-Net's strong capability in capturing true positive vessel pixels. DEU-Net [50] and Vessel-Net [8] also perform well with sensitivities of 80.74% and 81.32% respectively, indicating their effectiveness in detecting vessels. Next, Specificity evaluates the ability to correctly identify non-vessel pixels. CMD-Net excels with a specificity of 98.54%, indicating its superior

Table 9
Segmentation performance in Dice score of CMDC with different backbone models on six biomedical image datasets.

Method	CMDC	GlaS	CRAG	BISW	ESOS	DRIVE	CHASE
FCN-8s [13]	✗	79.7%	87.2%	70.1%	82.1%	95.7%	97.8%
	✓	86.0%	90.7%	75.1%	85.6%	97.3%	98.3%
U-Net [33]	✗	84.3%	88.6%	71.4%	85.7%	96.0%	98.1%
	✓	89.3%	91.3%	75.7%	88.4%	97.9%	98.8%
DeepLab-v3 [46]	✗	87.9%	91.3%	74.0%	86.2%	96.7%	97.9%
	✓	90.0%	93.1%	77.8%	89.1%	97.8%	98.5%
CMD-Net	✗	91.2%	92.0%	75.0%	89.5%	97.2%	98.3%
	✓	92.8%	93.5%	78.9%	91.2%	98.3%	98.9%

Table 10
Segmentation performance in mIoU score of CMDC with different backbone models on two natural image datasets.

Method	CMDC	VOC2012	Cityscapes
FCN-8s [13]	✗	62.2%	65.3%
	✓	66.1%	68.7%
U-Net [33]	✗	65.3%	66.8%
	✓	68.2%	69.3%
DeepLab-v3 [46]	✗	71.6%	63.1%
	✓	73.2%	65.9%
CMD-Net	✗	73.5%	70.7%
	✓	75.9%	73.3%

performance in avoiding false positives compared to HA-Net (98.13%), DEU-Net (98.21%), and Vessel-Net (98.14%). This metric highlights CMD-Net's ability to accurately delineate vessel boundaries while maintaining a low false positive rate. In terms of Accuracy, which reflects overall correctness in classification, CMD-Net achieves 96.93%. This places CMD-Net among the top performers alongside HA-Net (96.70%), DEU-Net (96.61%), and Vessel-Net (96.61%), indicating its robust performance in segmentation tasks on the CHASE-DB1 dataset. Lastly, In terms of AUC, CMD-Net achieves an AUC of 98.89%, demonstrating its strong discriminatory power similar to HA-Net (98.70%), and outperforming DEU-Net (98.12%). In summary, CMD-Net exhibits a balanced performance across Sensitivity, Specificity, Accuracy, and AUC on the CHASE-DB1 dataset, showcasing its effectiveness in retinal vessel segmentation.

The visual comparison of the vessel segmentation results is shown in Fig. 12. The images in the first two rows present the results on the DRIVE dataset, while the images in the last two rows show the results on the CHASEDB1 dataset. In the boxed portions, we can observe that some of the blood vessels to be segmented are extremely thin, with some regions being only 1–2 pixels wide, and the contrast is poor. Therefore, the segmentation of these challenging regions is highly demanding. Compared to the baseline method, the CMD-Net model can effectively improve the segmentation performance in these fuzzy areas. Overall, CMD-Net performs State-of-the-art performance on above four datasets, and its overall performance is competitive, demonstrating its effectiveness across different modalities and domain tasks. These results position CMD-Net as a promising model for advancing biomedical image segmentation research.

4.3.2. Generalization discussion

In this section, we add CMDC to a variety of backbones, and perform evaluation on six biomedical image datasets and two natural image datasets to demonstrate the generalization of CMDC in various domains. The backbone networks including FCN-8s [13], U-Net [33], and DeepLab-V3 [46]. Table 9 shows the comparison results on the adopted six biomedical image datasets. Overall, CMDC increases the performance of all four networks. For gland segmentation, an improvement of 3.8% and 2.4% is obtained on the GlaS dataset and the CRAG dataset, respectively. Furthermore, for vessel detection, our methods
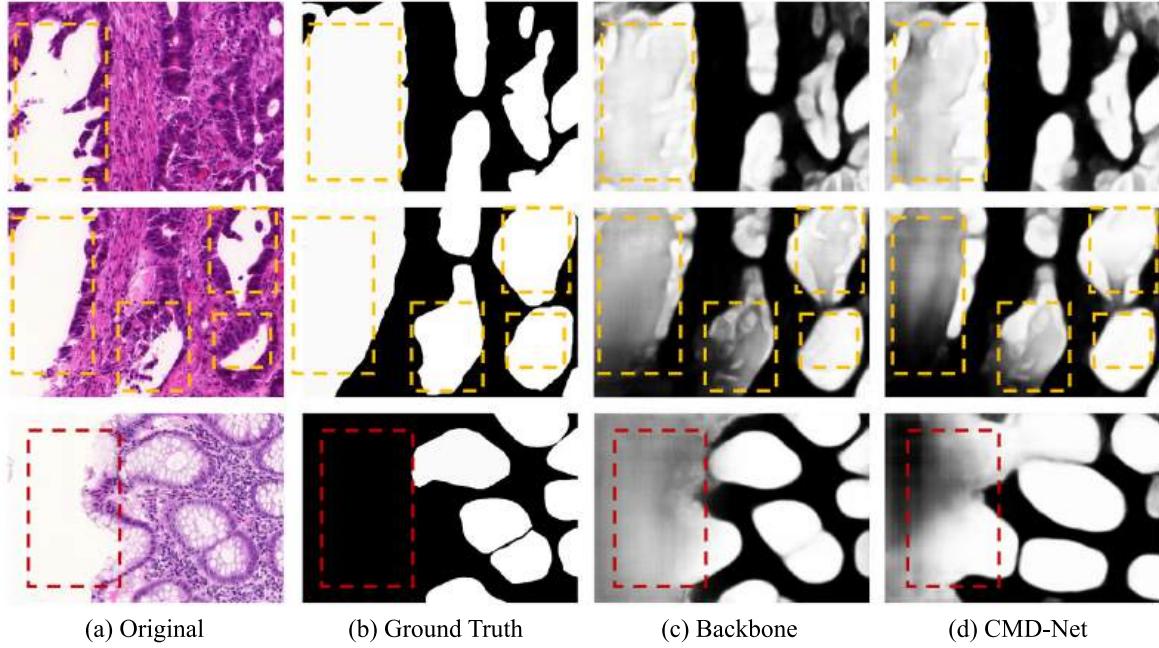
(a) Original   (b) Ground Truth   (c) Backbone   (d) CMD-Net

Fig. 13. Examples of hard cases in the GlaS dataset. Red and yellow rectangle boxes show the outside and inside parts of the ambiguous white areas, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

also achieve an improvement of 1.4% and 0.6% on the DRIVE dataset and the CHASE-DB1 dataset, respectively. Besides, on the BISW dataset and the ESOS dataset, our methods obtain an improvement of 4.3%, and 2.7%, respectively. Table 10 shows the comparison results on the adopted two natural image datasets. Similar experimental results observed that CMDC improve the all four networks by 1.6%–3.9%. In conclusion, the proposed CMDC can obviously improve a variety of networks on both biomedical image and natural image.

4.3.3. Limitation analysis

In fact, the limitation of CMD-Net mainly involves two aspects, which are analyzed as follow.

The first limitation is that improper image cropping leads to the limitation of the field of view of the training image itself on GlaS dataset. Fig. 13 depicts a number of challenging examples from the GlaS dataset. We can observe that the ambiguous white patches outside the gland and the gland lumen within the gland are more difficult to segment accurately. Beyond some boundary details, certain ambiguous white areas both within and outside the glands are poorly segmented. After consultation with clinical pathologists, we find that one possible cause is the mismatch between the object (gland) scale and the sampling scale in the source image. A typical gland has an inner tubular structure formed by a lumen and epithelial cell nuclei enclosing the cytoplasm. When the scale of the sampled patch and the gland tissue are nearly identical, it is common practice for primary pathologists to divide the gland tissue into two patches. However, if the two are mismatched with a significant difference, only a part of the object is included in the sampled image. Thus, without access to the original image, it is challenging for even the most qualified pathologists to identify the sampled area within such a limited receptive field. The first two rows of Fig. 13 depict typical examples of a poorly differentiated cancerous gland being sliced into two patches. We believe that the authors of the dataset have discovered this problem. CRAG [20] was published a few years after GlaS from the same author. The sample image size of CRAG (1512 × 1516) is about 6 times that of GlaS (775 × 522).

The second limitation is that only connecting feature with adjacent scales, alleviate the inconsistency of segmentation and the loss of local details, but limits the receptive field of feature learning. This problem

is more obvious in the gland segmentation domain. Because gland segmentation often requires the combination of global semantic clues to determine the specific category of local features. For example, the determination of some white areas mentioned in the previous section needs to be based on whether it is inside or outside the gland (global semantic clues) to determine whether it is a white area outside the tissue or an internal cavity. In fundus images, more local feature differences are used for judgment, so the impact is small. The most direct solution is to fuse high-dimensional semantic features (global semantic clues) with low-dimensional contour features (local contour details) through long connections, and then guide the learning of low-dimensional contour features. However, this long connection directly amplifies high-dimensional semantic features to align the scale with low-dimensional contour features, rather than amplifying scale by scale, which may also lead to incoherent segmentation. To address this limitation, we propose two future approaches. The first approach aims to further enhance effective feature sharing among scales. Our method demonstrates the existence of a specific connection pattern that improves performance and reduces computation cost compared to existing approaches. Therefore, future research can explore and expand on this idea. For example, Dense-Net may also possess certain connection patterns worth investigating. Furthermore, we suggest combining multi-scale feature aggregation methods with Dense-Net to explore useful connection patterns in a larger search space. The second approach involves enlarging the receptive field. Graph neural networks [52] are an effective method for expanding the receptive field. Thus, combining graph neural networks with U-Net may generate rich representations with a large receptive field. For instance, embedding a graph neural network into U-Net, similar to an attention module, may assist all layers in learning more effectively.

5. Conclusion

In this paper, we observe that current multi-scale dense connections are not optimal. To find the optimal connection configuration of dense connections, we introduce dense connection range, and adopt a naive approach and a network architecture search (NAS) based approach for discussion. we propose constrained multi-scale dense connections

(CMDC) by merely aggregating feature maps at the adjacent scales, including relevant features of the current scale. Experimental results showed that the CMDC-based CMD-Net can effectively improve the segmentation performance across a variety of networks and datasets with reduced computation cost, and achieves state-of-the-art performance on four public biomedical segmentation datasets.

CRediT authorship contribution statement

Jiawei Zhang: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Yanchun Zhang:** Supervision. **Hailong Qiu:** Data curation. **Tianchen Wang:** Writing – review & editing. **Xiaomeng Li:** Writing – review & editing. **Shanfeng Zhu:** Writing – review & editing. **Meiping Huang:** Supervision. **Jian Zhuang:** Supervision. **Yiyu Shi:** Supervision. **Xiaowei Xu:** Supervision.

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.

Data availability

Data will be made available on request.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT 3.5 in order to polish the manuscript. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Acknowledgments

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patcog.2024.111031.

References

[1] M. Hermsen, T. de Bel, M. Den Boer, E.J. Steenbergen, J. Kers, S. Florquin, J.J. Roelofs, M.D. Stegall, M.P. Alexander, B.H. Smith, et al., Deep learning–based histopathologic assessment of kidney tissue, J. Am. Soc. Nephrol. 30 (10) (2019) 1968–1979.

[2] M.Y. Ansari, Y. Yang, S. Balakrishnan, J. Abinahed, A. Al-Ansari, M. Warfa, O. Almokdad, A. Barah, A. Omer, A.V. Singh, et al., A lightweight neural network with multiscale feature enhancement for liver CT segmentation, Sci. Rep. 12 (1) (2022) 14153.

[3] K. Sirinukunwattana, J.P.W. Pluim, H. Chen, X. Qi, P.-A. Heng, Y.B. Guo, L.Y. Wang, B.J. Matuszewski, E. Bruni, U. Sanchez, N.M. Rajpoot, Gland segmentation in colon histology images: The glas challenge contest, MedIA 35 (2017) 489–502.

[4] H. Chen, X. Qi, L. Yu, P.-A. Heng, DCAN: deep contour-aware networks for accurate gland segmentation, in: CVPR, 2016, pp. 2487–2496.

[5] Z. Wen, J. Liu, Y. Li, GCSBA-net: Gabor-based and cascade squeeze bi-attention network for gland segmentation, IEEE J-BHI (2020).

[6] J. Staal, M.D. Abràmoff, M. Niemeijer, M.A. Viergever, B. Van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE TMI 23 (4) (2004) 501–509.

[7] H. Fu, Y. Xu, S. Lin, D.W.K. Wong, J. Liu, Deepvessel: Retinal vessel segmentation via deep learning and conditional random field, in: MICCAI, Springer, 2016, pp. 132–139.

[8] Y. Wu, Y. Xia, Y. Song, D. Zhang, D. Liu, C. Zhang, W. Cai, Vessel-net: retinal vessel segmentation under multi-path supervision, in: MICCAI, Springer, 2019, pp. 264–272.

[9] S. Li, J. Zhang, C. Ruan, Y. Zhang, Multi-stage attention-unet for wireless capsule endoscopy image bleeding area segmentation, in: BIBM, IEEE, 2019, pp. 818–825.

[10] Y. Fu, W. Zhang, M. Mandal, M.Q.-H. Meng, Computer-aided bleeding detection in WCE video, IEEE J-BHI 18 (2) (2013) 636–642.

[11] Y. Yuan, B. Li, M.Q.-H. Meng, Bleeding frame and region detection in the wireless capsule endoscopy video, IEEE J-BHI 20 (2) (2015) 624–630.

[12] X. Wang, Y. Chen, Y. Gao, H. Zhang, Z. Guan, Z. Dong, Y. Zheng, J. Jiang, H. Yang, L. Wang, et al., Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning, Nature Commun. 12 (1) (2021) 1–13.

[13] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: CVPR, 2015, pp. 3431–3440.

[14] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: CVPR, 2017, pp. 4700–4708.

[15] W. Ma, Y. Wu, F. Cen, G. Wang, Mdfn: Multi-scale deep feature learning network for object detection, PR 100 (2020) 107149.

[16] J. Wang, W. Zhang, J. Fu, K. Song, Q. Meng, L. Zhu, Y. Jin, Ldcnet: A lightweight multi-scale convolutional neural network using local dense connectivity for image recognition, TCDS (2023).

[17] L. Qi, J. Kuen, J. Gu, Z. Lin, Y. Wang, Y. Chen, Y. Li, J. Jia, Multi-scale aligned distillation for low-resolution detection, in: CVPR, 2021, pp. 14443–14453.

[18] L. Qian, C. Wen, Y. Li, Z. Hu, X. Zhou, X. Xia, S.-H. Kim, Multi-scale context unet-like network with redesigned skip connections for medical image segmentation, CMPB 243 (2024) 107885.

[19] J. Zhang, Y. Zhang, Y. Jin, J. Xu, X. Xu, Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation, HISC 11 (1) (2023) 13.

[20] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y.W. Tsang, N. Rajpoot, MILD-net: Minimal information loss dilated network for gland instance segmentation in colon histology images, MedIA 52 (2019) 199–211.

[21] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, T. Wang, A cross-modality learning approach for vessel segmentation in retinal images, IEEE TMI 35 (1) (2015) 109–118.

[22] S. Li, J. Zhang, Y. Jin, L. Zheng, J. Xu, G. Yu, Y. Zhang, Automatic segmentation of esophageal cancer pathological sections based on semantic segmentation, in: 2018 ICOT, IEEE, 2018, pp. 1–5.

[23] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, IJCV 88 (2010) 303–338.

[24] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: CVPR, 2016, pp. 3213–3223.

[25] C. Playout, R. Duval, F. Cheriet, A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images, in: MICCAI, Springer, 2018, pp. 101–108.

[26] P. Bilinski, V. Prisacariu, Dense decoder shortcut connections for single-pass semantic segmentation, in: CVPR, 2018, pp. 6596–6605.

[27] J. Chen, S. Banerjee, A. Grama, W.J. Scheirer, D.Z. Chen, Neuron segmentation using deep complete bipartite networks, in: MICCAI, Springer, 2017, pp. 21–29.

[28] 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, Springer, 2018, pp. 3–11.

[29] G. Li, M. Zhang, J. Li, F. Lv, G. Tong, Efficient densely connected convolutional neural networks, PR 109 (2021) 107610.

[30] X. Xiao, L. Wang, K. Ding, S. Xiang, C. Pan, Dense semantic embedding network for image captioning, PR 90 (2019) 285–296.

[31] J. Zhang, Y. Zhang, S. Zhu, X. Xu, Constrained multi-scale dense connections for accurate biomedical image segmentation, in: BIBM, IEEE, 2020, pp. 877–884.

[32] J. Fang, Y. Sun, Q. Zhang, Y. Li, W. Liu, X. Wang, Densely connected search space for more flexible neural architecture search, in: CVPR, 2020, pp. 10628–10637.

[33] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, Springer International Publishing, 2015, pp. 234–241.

[34] Y. Weng, T. Zhou, Y. Li, X. Qiu, NAS-unet: Neural architecture search for medical image segmentation, IEEE Access 7 (2019) 44247–44257.

[35] X. Yan, W. Jiang, Y. Shi, C. Zhuo, Ms-nas: Multi-scale neural architecture search for medical image segmentation, in: MICCAI, Springer, 2020, pp. 388–397.

[36] Z. Han, M. Jian, G.-G. Wang, ConvUNeXt: An efficient convolution neural network for medical image segmentation, KBS 253 (2022) 109512.

[37] M. Jafari, D. Auer, S. Francis, J. Garibaldi, X. Chen, DRU-net: an efficient deep convolutional neural network for medical image segmentation, in: ISBI, IEEE, 2020, pp. 1144–1148.

[38] M.Y. Ansari, Y. Yang, P.K. Meher, S.P. Dakua, Dense-PSP-UNet: A neural network for fast inference liver ultrasound segmentation, CBM 153 (2023) 106478.

[39] Y. Xie, J. Zhang, C. Shen, Y. Xia, Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation, in: MICCAI, Springer, 2021, pp. 171–180.

[40] L. Yang, Y. Zhang, J. Chen, S. Zhang, D.Z. Chen, Suggestive annotation: A deep active learning framework for biomedical image segmentation, in: MICCAI, Springer, 2017, pp. 399–407.

[41] J. Zhang, Y. Zhang, X. Xu, Pyramid u-net for retinal vessel segmentation, in: ICASSP, IEEE, 2021, pp. 1125–1129.

[42] Z. Yan, X. Yang, K.T. Cheng, A three-stage deep learning model for accurate retinal vessel segmentation, IEEE J-BHI 23 (4) (2019) 1427–1436.

[43] Y. Akhtar, S.P. Dakua, A. Abdalla, O.M. Aboumarzouk, M.Y. Ansari, J. Abinahed, M.S.M. Elakkad, A. Al-Ansari, Risk assessment of computer-aided diagnostic software for hepatic resection, IEEE TRPMS 6 (6) (2021) 667–677.

[44] M.Y. Ansari, A. Abdalla, M.Y. Ansari, M.I. Ansari, B. Malluhi, S. Mohanty, S. Mishra, S.S. Singh, J. Abinahed, A. Al-Ansari, et al., Practical utility of liver segmentation methods in clinical surgeries and interventions, BMC Med. Imaging 22 (1) (2022) 97.

[45] M.Y. Ansari, S. Mohanty, S.J. Mathew, S. Mishra, S.S. Singh, J. Abinahed, A. Al-Ansari, S.P. Dakua, Towards developing a lightweight neural network for liver ct segmentation, in: MICAD, Springer, 2022, pp. 27–35.

[46] L.-C. Chen, G. Papandreou, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE TPAMI 40 (4) (2018) 834–848.

[47] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A.L. Yuille, L. Fei-Fei, Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation, in: CVPR, 2019, pp. 82–92.

[48] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE TPAMI 39 (12) (2017) 2481–2495.

[49] S. Guo, K. Wang, H. Kang, Y. Zhang, Y. Gao, T. Li, BTS-dsn: Deeply supervised neural network with short connections for retinal vessel segmentation, IJMI 126 (2019) 105–113.

[50] B. Wang, S. Qiu, H. He, Dual encoding u-net for retinal vessel segmentation, in: MICCAI, Springer, 2019, pp. 84–92.

[51] D. Wang, A. Haytham, J. Pottenburgh, Y. Tao, Hard attention net for automatic retinal vessel segmentation, IEEE J-BHI (2020).

[52] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE TNN 20 (1) (2008) 61–80.

Jiawei Zhang received his Bachelor's degree from Xidian University in 2017 and his Ph.D. degree in Computer Science from Fudan University, Shanghai, China. Jiawei specializes in biomedical image analysis. He is an Assistant Professor in the Department of New Network at Peng Cheng Laboratory, Shenzhen, China.

Yanchun Zhang received his Ph.D. degree in Computer Science from the University of Queensland, Australia, in 1991. He is a Professor and the Director of the Centre for Applied Informatics at Victoria University, Melbourne, Australia.

Hailong Qiu received his B.S. degree in Clinical Medicine from Huazhong University of Science and Technology in 2016, and his M.S. degree in Clinical Medicine from South China University of Technology in 2019. He is currently pursuing his Ph.D. degree in Surgery at Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangzhou, China.
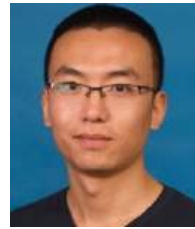
Tianchen Wang is currently pursuing his Ph.D. degree in Computer Science and Engineering at the University of Notre Dame. His research focuses on the applications and performance enhancement of deep learning and computer vision.

Xiaomeng Li is an Assistant Professor at The Hong Kong University of Science and Technology. Her research lies in the interdisciplinary areas of artificial intelligence and medical image analysis. Before joining HKUST, she was a Postdoctoral Research Fellow at Stanford University, working with Professor Lei Xing. She obtained her Ph.D. degree from The Chinese University of Hong Kong, advised by Professors Pheng-Ann Heng and Chi-Wing Fu.

Shanfeng Zhu received his B.S. and MPhil degrees in Computer Science from Wuhan University, Wuhan, China, in 1996 and 1999, respectively, and his Ph.D. degree in Computer Science from the City University of Hong Kong, Hong Kong, in 2003. He is a Professor in the School of Computer Science at Fudan University.

Meiping Huang received her M.S. degree in Imaging and Nuclear Medicine from Jinan University, Guangzhou, in 1999. She is currently a Professor of Medicine in the Department of Catheterization Lab at Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangzhou, China.

Jian Zhuang is currently a Professor of Medicine in the Department of Cardiac Surgery at Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangzhou, China.

Yiyu Shi received his Ph.D. degree in Electrical Engineering from the University of California at Los Angeles, Los Angeles, CA, USA, in 2009. He is currently a Professor in the Department of Computer Science and Engineering at the University of Notre Dame, Notre Dame, IN, USA.

Xiaowei Xu received his B.S. and Ph.D. degrees in Electronic Science and Technology from Huazhong University of Science and Technology, Wuhan, China, in 2011 and 2016, respectively. He is currently an Assistant Professor at Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangzhou, China.