Constrained Multi-scale Dense Connections for Accurate Biomedical Image Segmentation

Jiawei Zhang*†, Yanchun Zhang†‡, Shanfeng Zhu*, Xiaowei Xu§
*Shanghai key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China
Email: 17110240008, zhusf@fudan.edu.cn
†Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China
‡College of Engineering and Science, Victoria University, Melbourne, Australia
Email: yanchun.zhang@vu.edu.au
§ Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital, Guangzhou, China
Email: xiao.wei.xu@foxmail.com

***Abstract*—Biomedical image segmentation plays a critical role in clinical diagnosis and medical intervention. Recently, a variety of deep neural networks have boosted the biomedical image segmentation performance with a large margin, which adopts dense connections to explore rich representations in multiple scales. In multi-scale dense connections, features from all or most scales are fused or iteratively aggregated. In this paper, we propose constrained multi-scale dense connections (CMDC) for accurate biomedical image segmentation, which only fuse features from the nearest scales containing the most relevant appearance or semantic information. Based on CMDC, we further construct constraint multi-scale dense networks (CMD-Net) by applying CMDC to existing segmentation networks. Experiments across various architectures (including FCN-8s, U-Net, and DeepLabV3) and datasets (including GlaS, CRAG, KID, and ECS) demonstrate that CMD-Net not only outperforms existing schemes on both accuracy and efficiency but also can be easily generalized to a variety of segmentation networks. In addition, CMD-Net achieves state-of-the-art performance on two instance segmentation datasets, GlaS and CRAG.**

***Index Terms*—Biomedical Image Segmentation, Deep Learning, Constrained Multi-scale Dense Connections.**

I. Introduction

ACCURATE segmentation of biomedical issues such as the morphology of histological structures including colon [5], esophagi [6], and kidney [7] is an essential pre-requisite to obtain reliable morphological statistics, which is widely used for quantitative diagnosis. Conventionally, manual segmentation is performed by expert pathologists, which is laborious, time-consuming, and suffers from subjectivity among pathologists. Automatic segmentation methods are highly demanded in clinical practice to improve efficiency as well as reliability and reduce the workload of pathologists.

Recently, Fig. 2 illustrates many works [1]–[3] that have incorporated dense connections to make full use of multi-scale context information for performance improvement, which integrates all or most scales feature maps from the previous layers. Multi-scale dense connections can strengthen feature propagation and encourage feature reuse, which are implemented by small and large-scale transformations from long-range skip connections. However, as Fig. 1 shows, such transformations cannot preserve the original appearance information well from shallow layers. For example, due to various scaled glands, tiny gaps exist between adjacent glands and adhesive glands, which are rather difficult to separate from a complex background. A recent work [8] also observes a similar phenomenon. In addition, without step-by-step replenishing low-level information, these transformations usually lead to non-smooth segmentation results and sometimes even segmentation contradictions. This is particularly evident for tiny gaps between adjacent glands, which is smaller than the down-sampling scale (as indicated by the red dashed boxes in Fig. 1). From the above two observations, we find that some of the long-range connections in the dense connections may not contribute to or even harm the overall performance. We argue that proper connections to fuse feature maps in high and low resolutions may potentially further benefit the overall segmentation performance.

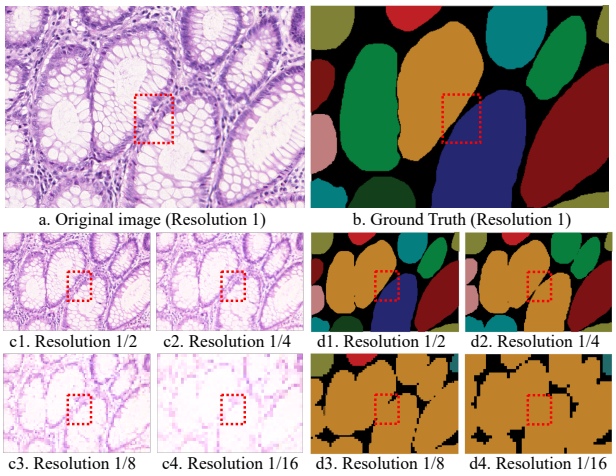


Fig. 1. Motivation illustration of the proposed constraint multi-scale dense connection (CMDC). Examples of (a) an original gland image, (b) its ground truth, and (c1-c4, d1-d4) their scaled results using max-pooling. Different colors represent different components. Small-scale transformations can preserve most appearance information, while large-scale transformations lose such critical information. Proper or constraint connections that only use small-scale transformations may potentially achieve higher segmentation performance.

978-1-7281-6215-7/20/$31.00

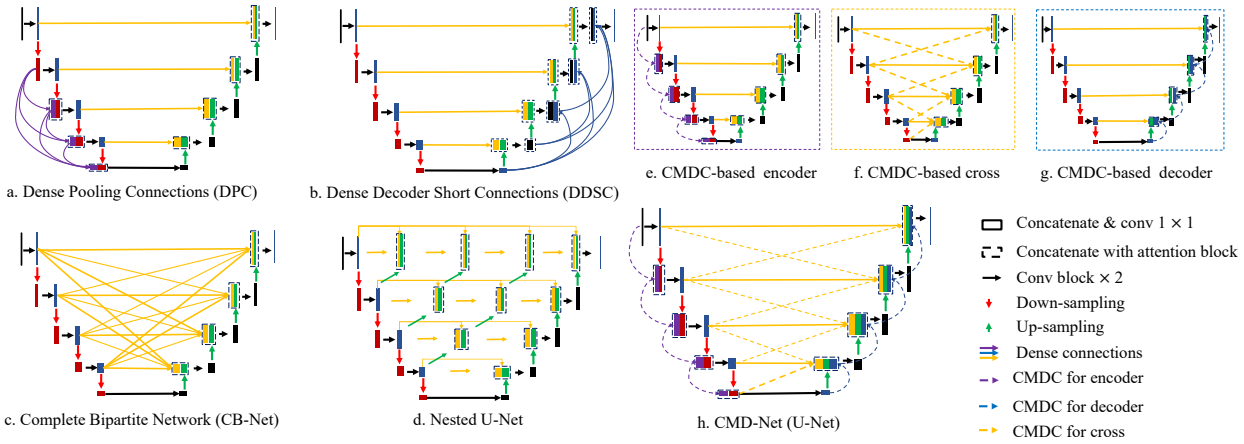


Fig. 2. Illustrations of (a-d) current multi-scale dense connections methods [1]–[3] and (e-h) our proposed constrained multi-scale dense connections (CMDC) method [4]. *CMDC-based encoder*, *CMDC-based decoder* and *CMDC-based cross* represent the corresponding structure by applying CMDC to the encoder, the decoder, and their cross, respectively. While CMD-Net is the combination of the above three.

Motivated by the above, in this paper, we propose constrained multi-scale dense connections (CMDC) that only fuse feature maps at nearest instead of all scales. We further applied CMDC to various positions (the encoder, decoder, and their cross) of segmentation networks, and propose constrained dense networks (CMD-Net) for accurate biomedical image instance segmentation. With CMDC, CMD-Net can capture morphological features in various scales to deepen the representations, reduce the semantic gaps between the encoder and decoder sub-networks and step-by-step use coarse-to-fine context information to refine resolution. We have conducted comprehensive experiments on various segmentation networks including FCN-8s [9], U-Net [10], DeepLab-V3 [11], SegNet [12], and four biomedical image segmentation datasets including gland image segmentation [5], [13], esophageal cancer image segmentation [6], and Wireless Capsule Endoscopy (WCE) image bleeding area segmentation [14]. The results show that our method can significantly improve the segmentation performance universally on all segmentation networks and datasets. Particularly, CMD-Net achieves state-of-the-art performance on GlaS and CRAG datasets. In addition, our method has lower computation complexity thus more efficient than existing works. In summary, our contributions are as follows:

- We introduce constrained multi-scale dense connections (CMDC) for accurate biomedical segmentation by only use feature maps at the nearest scales, which contains the most relevant feature from the original scale.
- Based on CMDC, we further propose constrained multi-scale dense networks (CMD-Net) by applying it to the encoder, decoder and their cross.
- Experiments cross architectures and tasks show that our constrained multi-scale dense connections can efficiently improve segmentation performance. Importantly, our proposed CMD-Net achieves state-of-the-art performance on GlaS and CRAG datasets.

We will briefly review related work in Section II. Section III introduces the proposed method, including CMDC and CMD-Net, in detail. The experimental setting, qualitative results, and ablation analysis are presented in Section IV, and the conclusion is given in Section V.

II. Related Work

A. Multi-scale Dense Connection

Recently, many works [15]–[21] adopted multi-scale dense connection to alleviate the vanishing-gradient problem, strengthen feature propagation, and encourage feature reuse. Such works can be divided into two main approaches: direct fusion of all scales and iterative fusion. The former allows information propagation from one block to another, as well as for multi-scale feature fusion. For example, dense pooling connections [1], complete bipartite networks [2] and dense decoder short connections [3] integrated full or most scales feature maps from previous layers on the encoder, decoder sub-network and cross them, respectively. The latter iteratively aggregates features at neighboring scales such as Nested U-Net [4], which drew lessons from deep layer aggregation [22], and connected the encoder and decoder through a series of nested, dense skip connections. Iterative fusion usually adds a large number of intermediate convolutions for performance improvement. Some other works [23]–[25] which employed dense connection within a single scale or a module without reusing feature maps at different scales, or maintained multi-scale representations concurrently, and these methods are out of the scope of this paper.

B. Biomedical Image Segmentation

In the light of U-Net [10], fully convolutional networks have dominated biomedical image segmentation such as gland image segmentation [26]–[29], esophageal cancer image segmentation [6], [30] and WCE image bleeding area segmentation [14], [31]. For instance, DCAN [32] designed a contour recognition decoding branch for unified multi-task learning

and the object boundaries were well segmented. To relieve the effort of manual annotations, active learning approaches are widely explored. Suggestive Annotation [33] selected the most representative samples actively based on uncertainty and similarity estimation. MILD-Net [13] introduced a rather deep structure by incorporating a minimal information loss unit to counter the loss of information in down-sampling.

III. Methods

In this section, we first introduce CMDC by applying CMDC to the encoder, decoder, and their cross, and then we fuse them to construct CMD-Net with examples on a variety of existing networks. For ease of explanation, we use U-Net as a vehicle in the following discussion. Note that CMDC can be combined with a variety of networks to form a series of CMD-Net.

A. Constrained Multi-scale Dense Connections

CMDC based Encoder/Decoder: The structure of a traditional encoder is illustrated in the dashed box in Fig. 3, and X^{i-1} and X^{i} are the input and output of the current layer. In order to integrate learned context information from neighboring scales to deepen and refine the representations, we use feature map X^{i-2} at the previous nearest scale to strengthen the input feature map. Particularly, we replace X_d^{i-1} with X_{newd}^{i-1}, which is composed of X_d^{i-1} and X_{new}^{i-2} (Eq. 1). Note that X_{new}^{i-2} is the down-sampling result of X^{i-2}, and $H(\cdot)$ denotes feature concatenation and 1×1 convolution, which is used to fuse channel-wise information and adjusts the number of channels as the same as that in the previous layer.

$$X_{newd}^{i-1} = H(X_{new}^{i-2}, X_d^{i-1}) \tag{1}$$

CMDC based decoder has a similar form as follows,

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

Where Y_{newp}^{i-1} is the strengthened results of Y_p^{i-1} by constrained dense decoder connections fusing the feature map at the previous scale in decoder for better feature presentation.

CMDC based Cross of Encoder and Decoder: CMDC based cross of the encoder and the decoder is shown in Fig. 4. Y^{i-1} and Y^{i} are the input and output of the current layer, respectively. Y_p^{i-1} is the intermediate output of Y^{i-1} after up-sampling. Traditionally, U-Net only fuses the same scale

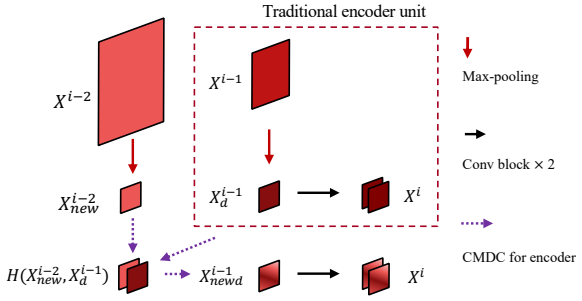


Fig. 3. Network structure of the CMDC-based encoder.

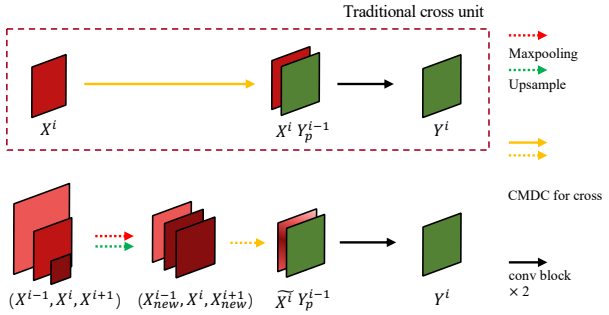


Fig. 4. Network structure of the CMDC-based cross of the encoder and the decoder.

feature maps X^{i} in the encoder to strengthen the feature maps Y^{i} in the decoder, while CMDC based encoder or decoder only fuses feature maps at the nearest scale with a larger resolution. Unlike the above two, CMDC based cross fuses feature maps from three scales: the one with the same scale, and two nearest scales with higher and lower resolutions. $\widetilde{X}^{i-1}$ is used to replace X^{i-1}, and $\widetilde{X}^{i-1}$ is computed as Eq. 3.

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

CMDC based cross integrates coarse-to-fine context information from the encoder sub-network to refine and deepen feature presentation. Note that CMDC based cross only replaces the original input at each scale by the fused coarse-scale features, and thus does not change the number of channels in the original convolution layer. Therefore, the added computation complexity is tiny, and the network structure modification is small, which makes it easy to be applied to other encoder-decoder based networks.

B. Constrained Multi-scale Dense Networks

CMD-Net is a combination of CMDC-based encoder, decoder, and cross of them. The individual implementation of each part has been discussed in the previous subsection, and the fusion of CMDC-based decoder and cross is detailed as follows:

$$\widetilde{Y}^{i-1} = H(Y_{newp}^{i-1}, \widetilde{X}^{i-1}, Y_p^{i-1}) \tag{4}$$

$$\Psi(\widetilde{\mathbf{Y}}^{\mathbf{i-1}}) = \sigma(\mathbf{W}(\mathbf{\Phi_{Avg}}(\widetilde{\mathbf{Y}}^{\mathbf{i-1}})) + \mathbf{W}(\mathbf{\Phi_{Max}}(\widetilde{\mathbf{Y}}^{\mathbf{i-1}}))) \otimes \widetilde{\mathbf{Y}}^{\mathbf{i-1}} \tag{5}$$

Where Y_{newp}^{i-1} and $\widetilde{X}^{i-1}$ are output feature maps of CMDC-based cross and decoder, respectively, Y_p^{i-1} is the output feature maps of the transpose operation in the decoder. $\Psi(\cdot)$ denotes the channel attention operation, which can enhance the integration by exploiting the inter-channel relationship of the features. Average-pooling $\mathbf{\Phi_{Avg}}(\cdot)$ and max-pooling $\mathbf{\Phi_{Max}}(\cdot)$ are adopted for aggregating channel information. $\sigma(\cdot)$ denotes the Sigmoid function. $\mathbf{W}(\cdot)$ is the 1*1 convolution and ReLU, while $\otimes$ denotes element-wise multiplication.

A series of CMD-Net can be obtained by applying the above modifications to existing works, as shown in Fig. 5. CMD-Net can easily take existing works such as FCN-8s [9], U-Net [10], and DeepLab-V3 [11] as backbones with minor modifications.

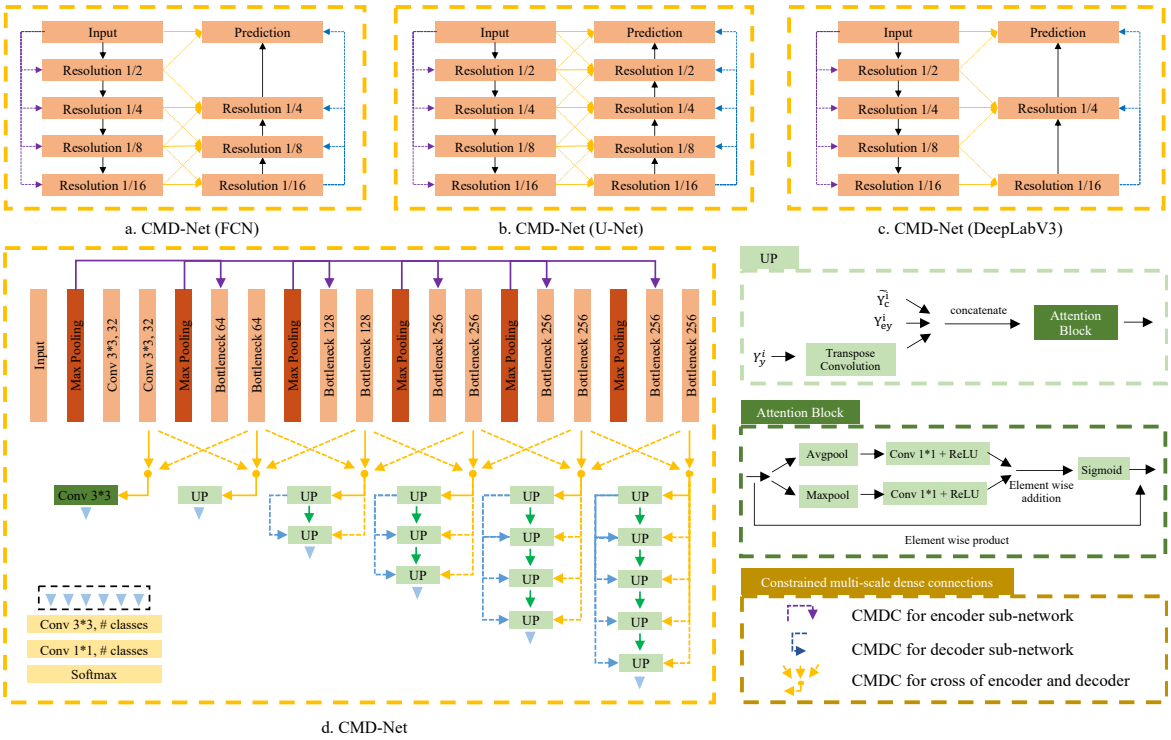


Fig. 5. Network structure of the proposed constrained multi-scale dense networks (CMD-Net). By applying CMDC to various segmentation networks including FCN-8s [9], U-Net [10], and DeepLab-V3 [11], a series of CMD-Net can be obtained. CMD-Net only enhances the input feature maps at each scales without massive modification of the internal module, which is convenient to apply our method to existing works. Note that purple, blue and yellow dotted lines represent CMDC in the encoder, decoder and their cross, respectively.

C. Implementation Details

More details of the implementation of our method, as shown in Fig. 5, are discussed as follows. In the **encoder** sub-network, the original convolution layers are replaced by residual modules with batch normalization, compared with traditional FCN, which can reduce the number of parameters while maintaining a similar number of feature channels at the end of each residual module. In the **decoder** sub-network, inspired by DCAN [32] and suggestive annotation [33], the structure is modified to gradually enlarge the size of the feature maps to ensure a smoothed result to cope with objects with various scales. A 3×3 convolution layer and a 1×1 convolution layer are applied to combine the feature maps from different branches together to better deal with glands of varied sizes. Besides, inspired by DCAN [32], an identical branch to detect contours is also used in our CMD-Net. The final result is the fusion of the probability values belonging to the object and the contour to boost the segmentation performance further.

IV. Experiments

A. Experimental Setup

*1) **Datasets**:* In the experiment, we used four medical imaging datasets for evaluation, including histology images and WCE images, to cover various medical imaging modalities.

- **GlaS**. The Gland Segmentation (GlaS) challenge held at MICCAI 2015 [5]. A Zeiss MIRAX MIDI slide scanner acquired the images from colorectal cancer tissues with a resolution of 0.62 μm/pixel. Images are mostly of size 775×522 pixels. They consist of a wide range of grades, from benign to malignant subjects. The dataset consists of 85 training (37 benign and 48 malignant) and 80 test images (37 benign and 43 malignant). Furthermore, the test images are split into an off-site set A and an on-site set B.
- **CRAG**. The Colorectal Adenocarcinoma Gland (CRAG) dataset origins from [13]. We have a total of 213 H&E CRA images taken from 38 WSIs, all of which are from different patients. Images are at 20 magnification and are mostly of size 1512 × 1516 pixels, with corresponding instance-level ground truth. The CRAG dataset is split into 173 training images and 40 test images with different cancer grades.
- **KID**. The KID dataset [14] is a database of high quality annotated Wireless Capsule Endoscopy (WCE) images.
- **ECS**. The Esophageal Cancer Segmentation (ECS) dataset [6] has 430 images, and the size is 830 × 1436 on average.

Note that the ground truth labels of GlaS and CRAG datasets are processed as instance-level annotation where a different integer marks every single gland instance.

*2) **Implementations**:* We trained our proposed methods on an NVIDIA GeForce GTX TITAN X (pascal), each containing 12 GB of memory. We adopted a batch size of 4 and set

TABLE I

INSTANCE SEGMENTATION COMPARISON OF CMD-NET WITH STATE-OF-THE-ART METHODS ON GLAS. **OBJECT F1 SCORE**, **OBJECT DICE** AND **OBJECT HAUSDORFF** EVALUATE THE **INSTANCE-LEVEL** PERFORMANCE OF DETECTION, SEGMENTATION AND SHAPE SIMILARITY RESPECTIVELY.

Method	Object F1-Score		Object Dice		Object Hausdorff	
	TestA	TestB	TestA	TestB	TestA	TestB
FCN-8s [9]	0.783	0.692	0.795	0.767	105.04	147.28
SegNet [12]	0.858	0.753	0.864	0.807	62.62	118.51
DeepLab-v3 [11]	0.862	0.764	0.859	0.804	65.72	124.97
DCAN [32]	0.912	0.716	0.897	0.781	45.42	160.35
Xu et al. (a) [28]	0.858	0.771	0.888	0.815	54.20	129.93
MIMO-Net [26]	0.913	0.724	0.906	0.785	49.15	133.98
Xu et al. (b) [29]	0.893	0.843	0.908	0.833	44.13	116.82
MILD-Net [13]	0.914	0.844	**0.913**	0.836	41.54	105.89
CMD-Net	**0.919**	**0.860**	0.912	**0.848**	**40.13**	**98.32**

the learning rate to 0.005 in the beginning. Besides, we choose Adam optimizer and cross-entropy loss to optimize the network. Furthermore, for fair comparisons, we used the same settings for all experiments. The same settings were also adopted in the ablation experiments. Also, we adopted the active learning method from [33] to train our CMD-Net. In the post-processing, since gland tissue is continuous and smooth, the raw segmentation output from CMD-Net was passed through a disk filter to smoother the segmented tissue mask, fill holes, and remove small areas. Thresholding was applied to restrict the pixels to binary values, and each connected component is labeled with a unique value for representing one segmented gland.

*3) **Evaluation**:* In experiment, Object F1-Score, Object Dice, and Object Hausdorff distance measurements were used to evaluate the instance-level detection, segmentation, and shape similarity performance of CMD-Net for gland image datasets (GlaS and CRAG). Particularly, the Object F1-Score is used in the assessment of gland instances detection. Different from traditional F1 score, the true positive is defined as the segmented object if at least 50% of it intersects with the ground truth, otherwise it's considered as a false positive. Dice metric can be formulated as $D(G,S) = 2|G \cap S|/(|G| + |S|)$, where G and S represent ground truth and prediction segmentation, respectively. Object Dice metric can be defined as $D_{object}(G,S) = \frac{1}{2}\left[\sum_{i=1}^{n_S} \omega_i D(G_i, S_i) + \sum_{j=1}^{n_G} \widetilde{\omega}_j D(\widetilde{G}_j, \widetilde{S}_j)\right]$ to evaluate instance-level segmentation results. $\omega_i = |S_i| / \sum_{m=1}^{n_S} |S_m|$, $\tilde{\omega}_j = \left|\tilde{G}_j\right| / \sum_{n=1}^{n_G} \left|\tilde{G}_n\right|$, n_S and n_G are the total number of segmented objects and ground truth objects, respectively. Finally, the shape similarity results are evaluated by Object Hausdorff metric, which is calculated as $H_{\text{object}}(G,S) = \frac{1}{2}\left[\sum_{i=1}^{n_S} \omega_i H\left(G_i, S_i\right) + \sum_{j=1}^{n_G} \tilde{\omega}_j H\left(\tilde{G}_j, \tilde{S}_j\right)\right]$, $H(G,S) = \max\left\{\sup_{x \in G} \inf_{y \in S} \|x-y\|, \sup_{y \in S} \inf_{x \in G} \|x-y\|\right\}$.

B. Results and Discussion

*1) **Comparison with State-of-the-art Works**:* The comparison of CMD-Net and existing works on GlaS and CRAG datasets are shown in Table I and Table II. Existing state-of-the-art methods include FCN-8s [9], DeepLab-v3 [11], SegNet [12], DCAN [32] and MILD-Net [13]. On Test A of GlaS, compared with the state-of-the-art method MILD-Net, CMD-Net obtains a higher performance on both Object F1-Score (improved by 0.5%) and Object Hausdorff (improved by 1.43), and a competitive result on Object Dice, which is only 0.1% worse. While on Test B of GlaS, CMD-Net achieves significantly better performance than MILD-Net, which is 1.6%, 1.2%, and 7.57 higher on Object F1-Score, Object Dice, and Object Hausdorff, respectively. It should be highlighted that Test B is much harder than Test A as it is with more malignant subjects. On CRAG dataset, CMD-Net outperforms the prior works by 1.5% on Object F1-Score, 0.4% on Object Dice, and 27.76 on Object Hausdorff.

TABLE II

INSTANCE SEGMENTATION PERFORMANCE COMPARISON OF CMD-NET WITH STATE-OF-THE-ART METHODS ON CRAG.

Method	Obj.F1	Obj.Dice	Obj.Hausdorff
FCN-8s [9]	0.558	0.640	436.43
SegNet [12]	0.622	0.739	247.84
DeepLab-v3 [11]	0.648	0.745	281.45
DCAN [32]	0.736	0.794	218.76
MILD-Net [13]	0.825	0.875	160.14
CMD-Net	**0.840**	**0.879**	**132.38**

The sample results of the comparison are shown in Fig. 6. We can notice that neighboring objects are extremely close in boxes marked region in Row 1 and 4. It's difficult to segment thees tiny gaps from those adjacent glands, e.g., the backbone method cannot divide the two objects in Row 1, and mistakenly segment a connected object into several small objects. Compared with existing works, CMD-Net can handle conditions well, especially at the object level. In addition, CMD-Net can also improve the segmentation of tiny glands (Row 3) and lumen with varied shapes (Row 2 and 5).

*2) **Ablation Discussion of CMDC Location**:* We discussed CMD-Net with different configurations of CMDC locations, and we also compared our methods with related multi-scale dense connections methods such as dense pooling connections (DPC) [1], CB-Net [2], Nested U-Net [4], and dense decoder short connections (DDSC) [3]. Ablation results on GlaS and CRAG datasets are shown in Table III. Compared

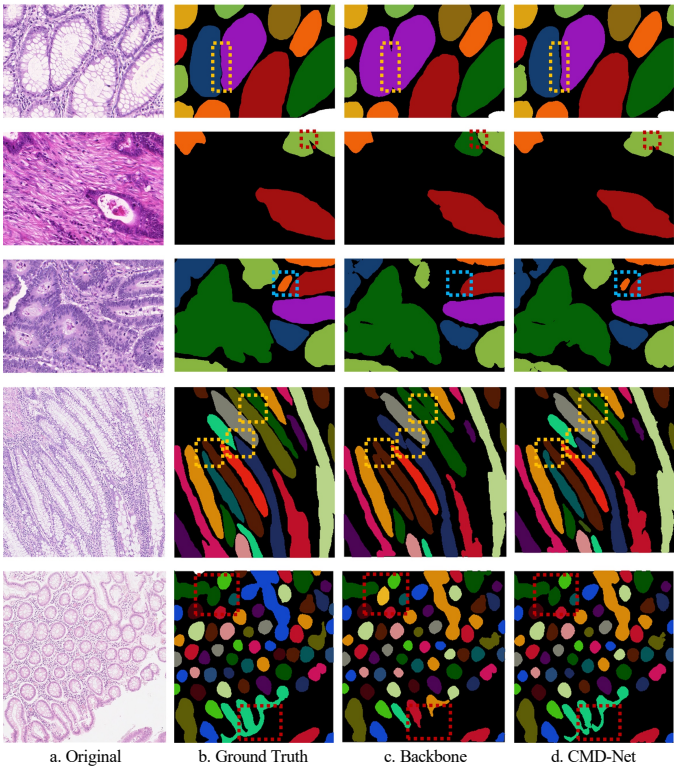


Fig. 6. Sample results of CMD-Net and existing works on GlaS (Row 1-3) and CRAG (Row 4-5) datasets. The improvements are indicated by rectangle boxes: yellow (false merge) , red (false split) and blue (false negative). Note that false merge has a greater influence on instance segmentation performance.

with corresponding multi-scale dense connections methods, our methods achieve an improvement of 1.7%, 1.2%, and 0.9% on average on the encoder, decoder sub-networks, and their cross, respectively. Compared with only applying CMDC to the encoder, decoder, or their cross, the combination of the three can obtain significantly higher performance, which is 2.15% on average. We can also find that CMDC-based encoder usually has a higher averaged improvement than CMDC-based decoder (0.85%) and cross (1.2%). It maybe caused by the fact that early feature sharing and reuse may have a broader impact on all the networks, thus improve the performance. In addition, our method consumes much fewer resources than existing works. Compared with U-Net, CMDC's memory, parameters, and FLOPs are increased by 1.7-40.2%, 0.4-5%, 19.3-12.2% on average, respectively. However, compared with existing methods, CMD-Net achieves 1.3-30%, 1.0-13.3%, and 127.0-80.0% lower consumption on memory, parameters, and FLOPs, respectively, while obtaining significantly higher performance. Even our CMD-Net consumes less memory, parameters, and FLOPs than Nested U-Net, which only implement multi-scale dense connections on the cross.

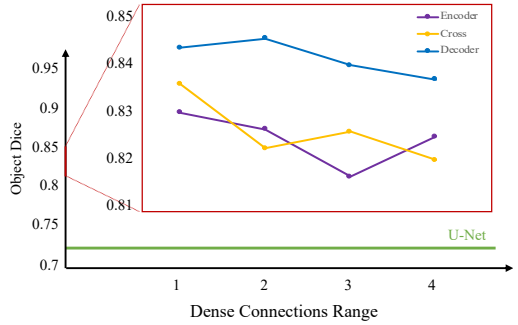


Fig. 7. Ablation experiment of dense connections range and locations on the part A of GlaS dataset. As the dense connections range increases, the accuracy fluctuates in a small interval, with no significant increase or even a slight decrease.

*3) **Ablation Discussion of Dense Connections Range**:* In the method section, we only take the nearest scales into consideration while ignores others. In fact, such a setup is supported by our ablation analysis, which is detailed in this part. For ease of discussion, we introduce ***dense connection range***, which is defined as the number of neighboring scales. For example, dense connection range 2 means each scale needs to consider five scales (two scales with higher resolutions, one scale with the same scale, and two scales with lower resolutions). The ablation results on GlaS dataset from the original U-Net and the one with different dense connection range on the encoder, decoder sub-networks, and their cross are shown in Fig. 7. Overall, CMDC can greatly improve the

TABLE III
ABLATION EXPERIMENTS OF CMDC LOCATIONS ON THE GLAS AND CRAG DATASETS. WE COMPARE THE **INSTANCE** SEGMENTATION PERFORMANCE ORIGINAL U-NET, U-NET WITH CMDC AND OTHER MULTI-SCALE DENSE CONNECTIONS METHODS.

Method	Location			Object F1-Score		Object Dice		Efficiency		
	E*	C	D	GlaS	CRAG	GlaS	CRAG	Memory(G)	Params(M)	FLOPs(G)
U-Net [10]				0.619	0.600	0.717	0.654	4.3	7.76	158.67
DPC [1]	✓			0.781	0.775	0.822	0.790	4.9	9.08	166.20
CMDC-based encoder	✓			**0.802**	**0.791**	**0.836**	**0.808**	**4.6**	**7.90**	**160.70**
Nested U-Net [4]		✓		0.733	0.741	0.792	0.753	10.4	9.05	391.72
CB-Net [2]		✓		0.760	0.778	0.818	**0.785**	-	**8.00**	198.75
CMDC-based cross		✓		**0.791**	**0.779**	**0.838**	0.781	**7.3**	8.07	**178.02**
DDSC [3]			✓	0.801	0.776	0.821	0.769	6.8	10.35	314.71
CMDC-based decoder			✓	**0.807**	**0.786**	**0.833**	**0.777**	**6.2**	**8.42**	**173.66**
CMDC U-Net	✓	✓	✓	**0.819**	**0.811**	**0.848**	**0.819**	**9.8**	**8.86**	**195.24**

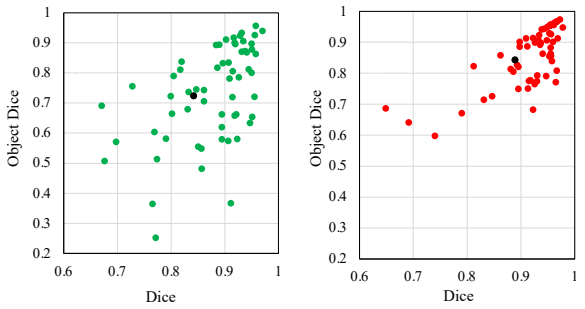
- means that can not train on a single GPU TITAN x with 12GB memory.
* E, C and D represent apply constrained multi-scale dense connections to encoder, decoder sub-network and cross them, respectively.

network instance segmentation performance, mostly by more than 10% on Object Dice. Meanwhile, we can find that as dense connection range increases, the accuracy fluctuates in a small interval, with no significant increase or even a slight decrease. This also shows that the network does not benefit from the fusion from too many scales.

TABLE IV
ABLATION EXPERIMENTS OF MAIN NETWORK STRUCTURES ON FOUR BIOMEDICAL IMAGE DATASETS. **DICE** EVALUATES THE **SEMANTIC-LEVEL** SEGMENTATION PERFORMANCE.

Methods	CMDC	GlaS		CRAG	KID	ESC
		TestA	TestB			
FCN-8s [9]	×	0.794	0.805	0.872	0.701	0.821
	✓	0.859	0.864	0.907	0.751	0.856
U-Net [10]	×	0.847	0.832	0.886	0.714	0.857
	✓	0.902	0.865	0.913	0.757	0.884
DeepLab-v3 [11]	×	0.889	0.847	0.913	0.740	0.862
	✓	0.911	0.868	0.931	0.778	0.891
CMD-Net	×	0.916	0.898	0.920	0.750	0.895
	✓	0.933	0.912	0.935	0.789	0.912

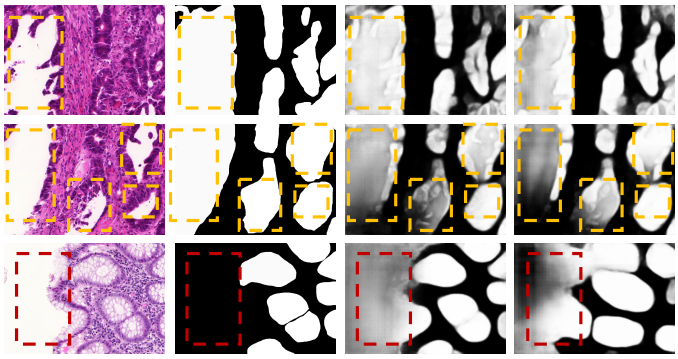
*4) **Ablation Discussion of Main Network Structure**:* To demonstrate the generalization of CMD-Net, we apply it to various semantic segmentation networks including FCN-8s [9], U-Net [10], and DeepLab-V3 [11]. The results on four datasets are shown in Table IV. Overall, CMD-Net improves the segmentation performance of all four networks. Particularly, improvement of 4.0% and 3.2% are obtained on Test A and Test B of GlaS on average, respectively. Meanwhile, our methods achieve 2.4%, 4.3% and 2.7% on CRAG, KID, ESC dataset, respectively. We can observe that simple structure can obtain a large improvement from our CMDC, while the methods without multi-scale information fused can achieve a better enhancement than those with.



a. Backbone b. CMD-Net

Fig. 8. Dice and Object Dice scores of U-Net (a) and U-Net with CMDC (b) on the part A of GlaS. The black dot stands for the average point. We can see that the instance segmentation of U-Net with CMDC has been greatly improved. This demonstrates that our CMDC greatly prevents the instance segmentation performance decreasing due to incorrect segmentation of details.

*5) **Impact on Instance Segmentation**:* Instance segmentation is more challenging than segmentation, and in this part, we further explore the impact of CMD-Net on instance segmentation. U-Net is adopted as the main network structure, and GlaS dataset is used. As shown in Fig. 8, obviously, Object Dice tends to be smaller than Dice if the neighboring cells are not well-segmented apart, indicating that the network fails to detect some challenging and significant ambiguous areas such as adjacent objects with tiny gaps. In Fig. 8(a), the distribution spreads across the whole plane, indicating that there are some results with good segmentation scores but with a bad performance on instance-level segmentation. In Fig. 8(b), CMDC improve the Dice and Object Dice scores simultaneously. The black dot presents the average performance. We can see that an average improvement of 4% is achieved on Dice. Particularly, a large margin of 12% on Object Dice is gained. Thus, CMD-Net can obtain a higher improvement on Object Dice than Dice, which shows that CMD-Net is especially effective in processing object segmentation.



a. Original b. Ground Truth c. Backbone d. CMD-Net

Fig. 9. Several hard cases on GlaS dataset. The columns represent the original image, ground truth, segmentation results of FCNsa and CMD-Net. The ambiguous white areas are indicated by rectangle boxes: red (outside) and yellow (inside).

*6) **Error Analysis of Segmentation Results**:* For ease of discussion, we focus the analysis on the gland segmentation of GlaS and CRAG datasets. By checking the results with low Dice scores, we find that the ambiguous white areas outside the gland and the gland lumen inside the gland are generally more difficult to segment. Fig. 9 shows several hard cases on GlaS dataset for discussion. We can notice that besides some boundary details, some ambiguous white areas inside and outside of the glands are segmented poorly. A potential reason is the improper sampling from the original image. Usually a typical gland comprises a lumen area forming the interior tubular structure and an epithelial cell nuclei surrounding the cytoplasm. When the sampled patch and the gland tissue are roughly the same scale, it is a general case to cut the gland tissue into two patches. In this way, even the experienced pathologist can not directly diagnose the sampled patch under limit receptive field without the original image. The images shown in Row 1-2 of Fig. 9 are a typical case where a poorly-differentiated malignant gland is cut into two patches. A possible solution is to enhance feature sharing and performs object level classification using graph neural network [34] in the largest scale (the lowest resolution).

V. CONCLUSION

In this paper, we proposed CMDC for accurate biomedical segmentation by only reuse feature maps at the nearest

scales. We further proposed a series of CMD-Net by applying CMDC to various locations of existing segmentation networks. Comprehensive experiments were conducted on four networks and four datasets, and the experimental results showed that CMDC can effectively improve the segmentation performance with reduced resource consumption. Furthermore, CMD-Net obtained state-of-the-art performance on two instance segmentation datasets, GlaS and CRAG. CMD-Net can be easily generalized to existing works with minor modifications.

REFERENCES

[1] C. Playout, R. Duval, and F. Cheriet, "A multitask learning architecture for simultaneous segmentation of bright and red lesions in fundus images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 101–108.

[2] J. Chen, S. Banerjee, A. Grama, W. J. Scheirer, and D. Z. Chen, "Neuron segmentation using deep complete bipartite networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 21–29.

[3] P. Bilinski and V. Prisacariu, "Dense decoder shortcut connections for single-pass semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6596–6605.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and 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.

[5] 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, A. Böhm, O. Ronneberger, B. B. Cheikh, D. Racoceanu, P. Kainz, M. Pfeiffer, M. Urschler, D. R. J. Snead, and N. M. Rajpoot, "Gland segmentation in colon histology images: The glas challenge contest," *Medical image analysis*, vol. 35, pp. 489–502, 2017.

[6] S. Li, J. Zhang, Y. Jin, L. Zheng, J. Xu, G. Yu, and Y. Zhang, "Automatic segmentation of esophageal cancer pathological sections based on semantic segmentation," in *2018 International Conference on Orange Technologies (ICOT)*. IEEE, 2018, pp. 1–5.

[7] 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," *Journal of the American Society of Nephrology*, vol. 30, no. 10, pp. 1968–1979, 2019.

[8] Z. Yan, X. Yang, and K.-T. T. Cheng, "A deep model with shape-preserving loss for gland instance segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 138–146.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[10] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Springer International Publishing, 2015.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[13] S. Graham, H. Chen, J. Gamper, Q. Dou, P.-A. Heng, D. Snead, Y. W. Tsang, and N. Rajpoot, "Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images," *Medical image analysis*, vol. 52, pp. 199–211, 2019.

[14] S. Li, J. Zhang, C. Ruan, and Y. Zhang, "Multi-stage attention-unet for wireless capsule endoscopy image bleeding area segmentation," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 818–825.

[15] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 179–187.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6230–6239.

[17] J. Zhang, Y. Jin, J. Xu, X. Xu, and Y. Zhang, "Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation," *arXiv preprint arXiv:1812.00352*, 2018.

[18] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[19] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected u-nets for efficient landmark localization," in *European Conference on Computer Vision (ECCV)*, 2018.

[20] G. Lin, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation." in *Cvpr*, vol. 1, no. 2, 2017, p. 5.

[21] L. Dong, L. He, M. Mao, G. Kong, X. Wu, Q. Zhang, X. Cao, and E. Izquierdo, "Cunet: a compact unsupervised network for image classification," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2012–2021, 2018.

[22] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2403–2412.

[23] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–19.

[24] S. Guan, A. Khan, S. Sikdar, and P. V. Chitnis, "Fully dense unet for 2d sparse photoacoustic tomography artifact removal," 2018.

[25] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *arXiv preprint arXiv:1902.09212*, 2019.

[26] S. E. A. Raza, L. Cheung, D. Epstein, S. Pelengaris, M. Khan, and N. M. Rajpoot, "Mimo-net: A multi-input multi-output convolutional neural network for cell segmentation in fluorescence microscopy images," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, April 2017, pp. 337–340.

[27] M. P. Shah, S. Merchant, and S. P. Awate, "Ms-net: Mixed-supervision fully-convolutional networks for full-resolution segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 379–387.

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

[29] Y. Xu, Y. Li, M. Liu, Y. Wang, Y. Fan, M. Lai, E. I. Chang *et al.*, "Gland instance segmentation by deep multichannel neural networks," *arXiv preprint arXiv:1607.04889*, 2016.

[30] D. Jin, D. Guo, T.-Y. Ho, A. P. Harrison, J. Xiao, C.-k. Tseng, and L. Lu, "Deep esophageal clinical target volume delineation using encoded 3d spatial context of tumors, lymph nodes, and organs at risk," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 603–612.

[31] X. Jia and M. Q.-H. Meng, "A study on automated segmentation of blood regions in wireless capsule endoscopy images using fully convolutional networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 179–182.

[32] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "Dcan: deep contour-aware networks for accurate gland segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 2487–2496.

[33] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 399–407.

[34] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.