Medical Image Analysis 82 (2022) 102626

Contents lists available at ScienceDirect

Medical Image Analysis

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

Semantic instance segmentation with discriminative deep supervision for medical images

Sihang Zhou [a], Dong Nie [b], Ehsan Adeli [c,d], Qian Wei [e], Xuhua Ren [f], Xinwang Liu [g,*], En Zhu [g,*], Jianping Yin [h], Qian Wang [i], Dinggang Shen [j,k,**]

[a] College of Intelligence Science and Technology, National University of Defense Technology, No. 109 Deya Road, Changsha, Hunan 410073, China
[b] Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3175, USA
[c] Department of Computer Science, Stanford University, Stanford, CA 94305, USA
[d] Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA
[e] Nanchang University Queen Mary School, Nanchang University, No. 1299 Xuefu Avenue, Nanchang city, Jiangxi Province 330000, China
[f] Tencent Technology, Hongmei Road 1801 Tencent building, Shanghai 200233, China
[g] School of Computer, National University of Defense Technology, No. 109 Deya Road, Changsha, Hunan 410073, China
[h] School of Cyberspace Science, Dongguan University of Technology, Guangdong 523808, China
[i] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China
[j] School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China
[k] Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200232, China

ARTICLE INFO

Keywords:
Semantic instance segmentation
Computer-aided diagnosis and therapy
Discriminative deep supervision
Semantic enhanced boundary detection

ABSTRACT

Semantic instance segmentation is crucial for many medical image analysis applications, including computational pathology and automated radiation therapy. Existing methods for this task can be roughly classified into two categories: (1) proposal-based methods and (2) proposal-free methods. However, in medical images, the irregular shape-variations and crowding instances (e.g., nuclei and cells) make it hard for the proposal-based methods to achieve robust instance localization. On the other hand, ambiguous boundaries caused by the low-contrast nature of medical images (e.g., CT images) challenge the accuracy of the proposal-free methods. To tackle these issues, we propose a proposal-free segmentation network with discriminative deep supervision (DDS), which at the same time allows us to gain the power of the proposal-based method. The DDS module is interleaved with a carefully designed proposal-free segmentation backbone in our network. Consequently, the features learned by the backbone network become more sensitive to instance localization. Also, with the proposed DDS module, robust pixel-wise instance-level cues (especially structural information) are introduced for semantic segmentation. Extensive experiments on three datasets, i.e., a nuclei dataset, a pelvic CT image dataset, and a synthetic dataset, demonstrate the superior performance of the proposed algorithm compared to the previous works.

1. Introduction

Semantic instance segmentation is a challenging task in computer vision since it requires both precise localization and accurate labeling of each instance in a given image (Pinheiro et al., 2015; Dai et al., 2016a,b). Segmenting the instances is crucial in many computer-aided medical image analysis applications, including computational pathology (Chen et al., 2016; Xu et al., 2017) and automated radiation therapy (Lessmann et al., 2018; Wang et al., 2019). However, for medical image segmentation, this task becomes even more complicated due to: (1) natural low-tissue-contrast and the artifacts embodied in medical images (like CT and MR images); (2) large scale, shape, and appearance variation of the target instances (e.g., nuclei and body organs); and (3) highly complex scenes (e.g., crowded nuclei in a single image). For more explicit illustration, please check Fig. 1. These factors result in complex cases with ambiguous boundaries to determine. The existing methods for this problem can be roughly categorized into two main categories: (1) proposal-based methods (like (Dai et al., 2016a,b,c; Li et al., 2017; He et al., 2017; Pinheiro et al., 2015; Liu et al., 2018)); and (2) proposal-free methods (such as (Bai and Urtasun, 2017; Chen et al., 2016; Xu et al., 2017; Salvador et al., 2017; De Brabandere et al., 2017; Kirillov et al., 2017; Huang et al., 2021)).

* Corresponding authors.
** Corresponding author at: Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200232, China.
E-mail addresses: xinwangliu@nudt.edu.cn (X. Liu), enzhu@nudt.edu.cn (E. Zhu), dgshen@shanghaitech.edu.cn (D. Shen).

https://doi.org/10.1016/j.media.2022.102626
Received 7 October 2021; Received in revised form 21 August 2022; Accepted 10 September 2022
Available online 24 September 2022
1361-8415/© 2022 Elsevier B.V. All rights reserved.

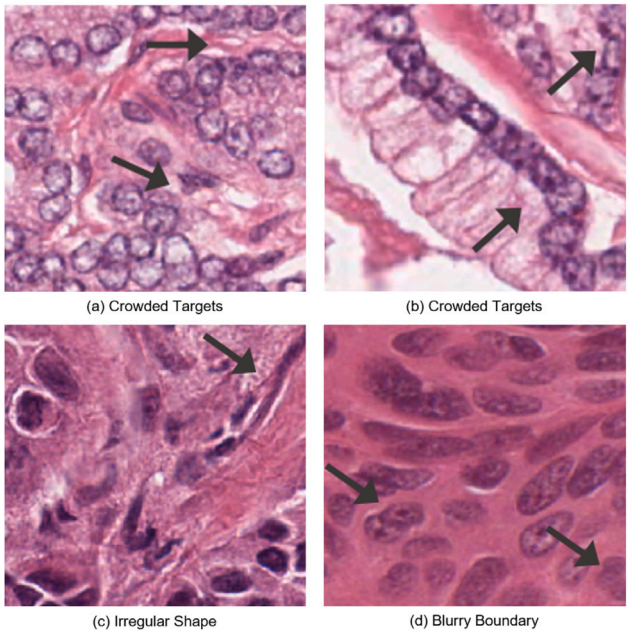


Fig. 1. Illustration of typical medical images. Here, nuclei images are adopted as an example. As can be seen from the figures, the crowded targets, irregular shapes, and blurry nuclei boundaries make the target localization an arduous task in this circumstance.

The *proposal-based methods* often adopt a typical object detection pipeline, in which the bounding boxes of the target instances are first detected and then refined to obtain the segmentation masks for instances (See Fig. 2a). Thanks to the detection operation integrated into this framework, the proposal-based methods can model structural and appearance variations of individuals in detail, thus capable of dealing with problems caused by occlusion and low image contrast (He et al., 2017). Consequently, these methods have achieved state-of-the-art performance in important natural image instance segmentation challenges like MSCOCO (Lin et al., 2014) and Pascal VOC2012 (Everingham et al., 2010). However, on the other hand, in these bounding box-centered segmentation methods, the quality of the generated bounding box proposals may pose a limit to the performance of these methods. Specifically, since these methods are based on a limited number of region proposals with predefined shapes, the robustness of these methods can be easily undermined by the crowded scenes, the inappropriately-set bounding box sizes and ratios, or the target instances with fundamentally non-rectangular shapes (Kirillov et al., 2017; Novotny et al., 2018). These problems are usually encountered in medical image instance segmentation. For example, in instance segmentation of microscopic images, the cell instances can be crowded with a large variety in shape and size (Chen et al., 2016).

The *proposal-free methods* usually start with a global foreground/background segmentation and then partition the foreground into target instances by exploiting instance-level clues such as instance boundaries (Kirillov et al., 2017), the interior of instances (Bai and Urtasun, 2017) (see Fig. 2b). Without the limitation of bounding box quality, the mechanism of the proposal-free methods are more suitable for the segmentation circumstances with crowded scenes and the large shape or size variation of the target instances to their proposal-based counterparts. Therefore, these methods have achieved promising performance in many important medical image segmentation competitions.[1] However, since in this branch of methods, instead of modeling the targets directly, the relatively low-level instance information (i.e., point-wise similarity (Novotny et al., 2018; De Brabandere et al., 2017; Fathi et al., 2017), instance boundaries (Chen et al., 2016; Kirillov et al., 2017; Liu et al., 2017), key points (Papandreou et al., 2018), and instance-related direction field (Bai and Urtasun, 2017; Kendall et al., 2018), etc.) is modeled for instance localization, more robust instance structural and shape information is insufficiently exploited (please check Section 2.2 for details). As a result, the performance of these methods is usually unstable on the low-contrast images, in which instance boundaries are ambiguous compared with textures from both foreground and background.

In this paper, we propose a novel proposal-free fully convolutional network (FCN) to combine the advantages of the two types of methods. The main idea of our algorithm is to endow the proposal-free instance segmentation algorithms with the ability to exploit the structural information of the target instances. In this way, the proposed algorithm would achieve preferable segmentation accuracy against not only crowded, irregular but also low contrast and ambiguous targets. This goal is achieved by introducing a detection-based structural information exploitation module to a proposal-free instance segmentation framework. Specifically, in our method, we adopt a multi-task segmentation framework to jointly learn the semantic segmentation maps, the boundary maps between instances, as well as the instance interior maps (Roth et al., 2018) (see Fig. 2c). A watershed algorithm is followed to generate the instance segmentation maps with these maps. Especially, to make the multi-task FCNs more discriminative to the vague and touching targets, we design a novel **detection-based deep supervision (DDS) module**. In this module, we conduct instance detection using the feature maps of the three sub-tasks to inject instance-sensitive structural information to the backbone FCN intrinsically. Moreover, we concatenate the feature maps learned by the detection sub-task with those in the segmentation branches. Since in the detection feature maps, each feature vector is used to predict whether a target appears in the bounding boxes centered at a pixel, these feature vectors collect the overall structural, shape, and appearance information of an instance (see Fig. 2c). This kind of information is essentially different from that provided by the low-level tasks such as boundary detection, which only tells whether a specific pixel is on the instance boundary or not. Thanks to this setting, we can introduce **pixel-wise instance-level information** into the FCN network for better instance detection and segmentation performance through the concatenation operation. Superior experimental results on three datasets, i.e., a nuclei dataset, a pelvic CT image dataset, and a synthetic dataset, show the effectiveness of our proposed discriminative deep supervision-based FCN in solving various segmentation challenges. The results and illustrations suggest that with the enhancement of the proposed DDS module, the learned features can better capture structural information of instances and are thus more discriminative for the blurry adjacent instances.

In summary, the novelties of this paper are three-fold. (1) We propose a novel and effective instance segmentation network to alleviate the intrinsic drawbacks of both proposal-free and proposal-based methods on segmenting blurry or crowded medical images. (2) We integrate a detection task with the backbone segmentation network using a deep supervision scheme for robust instance-discriminative information induction. (3) Instead of using the detection module for extracting ROIs (as usually done in previous methods), we concatenate the feature maps of the module with those in the backbone segmentation network, thus introducing robust pixel-wise instance-level information to our proposal-free segmentation network.

2. Related work

This section introduces the highly related proposal-based methods, proposal-free methods, and deep supervision methods in detail. The differences between our work and these methods are also carefully analyzed.

[1] https://www.kaggle.com/c/data-science-bowl-2018.

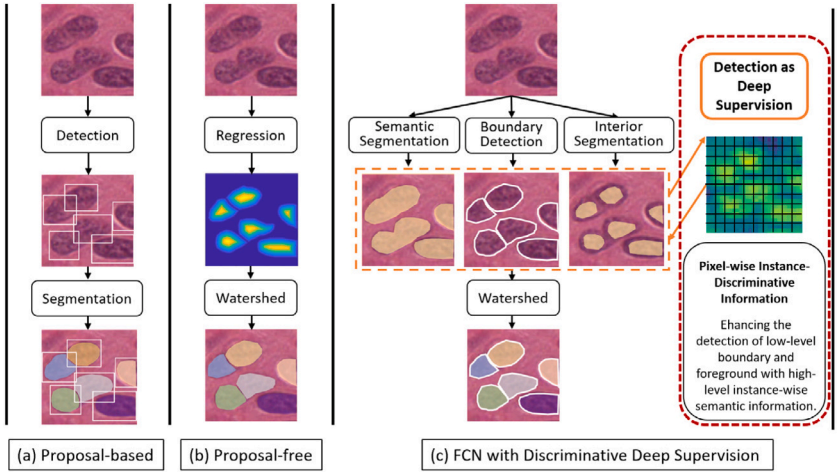


Fig. 2. The pipeline of three typical methods: (a) proposal-based method (i.e., Multitask Network Cascades (Dai et al., 2016b)), (b) proposal-free method (i.e., deep watershed transform (Bai and Urtasun, 2017)), and (c) our proposed discriminative deep supervision-based FCN (DDS-FCN). Since, with the proposed DDS module, our model can introduce discriminative pixel-wise instance-level information to the FCN, our model can work robustly against the limitation of detection accuracy while having more sensitivity to the blurry and complex touching boundaries. Consequently, our algorithm can perform better in medical image instance segmentation scenarios even with densely-distributed targets with significant shape variations.

2.1. Proposal-based instance segmentation

Early proposal-based instance segmentation methods integrated two sub-networks for object detection and segmentation separately and sequentially (Dai et al., 2016a,b; Zagoruyko et al., 2016), leading to immense computational complexities and often sub-optimal solutions (Li et al., 2017). To reduce computational time and make full use of the convolutional feature maps, (Li et al., 2017) proposed the first fully convolutional end-to-end solution to jointly detect and segment the object instances. To further improve the performance of the instance segmentation, parallel network structures (that can conduct segmentation and detection simultaneously) are commonly adopted by recent methods (Chen et al., 2017; Arnab and Torr, 2017; He et al., 2017; Liu et al., 2018). Among these methods, Mask R-CNN-based algorithms (He et al., 2017; Liu et al., 2018) are currently the leading frameworks in several benchmarks. Although great performance has been achieved by this type of methods, the bounding box-oriented segmentation mechanism limits these methods by the quality of the adopted detector since the objects missed by the detector will never have a chance to be correctly segmented. In contrast, although bounding boxes are also generated as a byproduct in our proposed proposal-free method and can be utilized to further improve the segmentation accuracy, the detection module is only loosely integrated to provide robust instance-level information for the three sub-networks. As a consequence, our network enjoys the discriminative instance recognition capacity of the proposal-based methods but without being influenced by the restrictions of the bounding box proposal mechanism.

2.2. Proposal-free instance segmentation

To segment instances, two strategies are usually adopted by the proposal-free methods. The first one generally adopts two-stage processing. In this strategy, researchers first conduct pixel-level predictions by semantic segmentation modules and then adopt a clustering process to group the sub-regions for each object instance (Novotny et al., 2018; De Brabandere et al., 2017; Fathi et al., 2017). The second strategy adopts a multi-task learning mechanism and incorporates the semantic segmentation with instance-related tasks, i.e., boundary detection (Chen et al., 2016; Kirillov et al., 2017), instance-level direction field regression (Bai and Urtasun, 2017; Kendall et al., 2018), object breakpoints (Liu et al., 2017), and key points (Papandreou et al., 2018) detection. With the help of the instance-related clues, semantic segmentation maps are further refined to fit each instance. Our proposed method falls into the second category. The mentioned algorithms have achieved promising performance in many applications. However, in the segmentation target functions of most existing methods in both strategies, a pixel as a unit is modeled. Taking the boundary detection task (Chen et al., 2016; Kirillov et al., 2017; Liu et al., 2017; Huang et al., 2021) as an example, the probability of each pixel being on the boundary of an instance is learned:

$$f(\mathbf{x}_{i,j}, \boldsymbol{\theta}) = \mathbf{y}_{i,j},\ \mathbf{y}_{i,j} \in \{0,1\}_{1\times 2}. \tag{1}$$

In Eq. (1), $\mathbf{x}_{i,j}$ and $\mathbf{y}_{i,j}$ are the feature vector and the one-hot label of pixel $\mathbf{v}_{i,j}$, respectively. $f(\cdot, \boldsymbol{\theta})$ is the network and the corresponding parameters. In this circumstance, a pixel as a unit is modeled. The network learns the probability of each pixel belongs to instance boundary or not. Comparatively, in our proposed discriminative deep supervision module, a bounding box as a unit is modeled. For example, for the classification sub-task in this module, the class of the bounding box represented by the feature vector $\mathbf{x}_{i,j}$ is predicted:

$$f^*(\mathbf{x}^*_{i,j}, \boldsymbol{\theta}^*) = \mathbf{y}^*_{i,j},\ \mathbf{y}_{i,j} \in \{0,1\}_{k\times c}. \tag{2}$$

In Eq. (2), k is the per-pixel anchor box number (Ren et al., 2015; He et al., 2017), and $\mathbf{y}^*_{i,j}$ denotes the class of the anchor boxes represented by feature vector $\mathbf{x}^*_{i,j}$. As a consequence, the integrated DDS module treats the target instance as a whole and introduces more robust instance-level information to the network. The difference between Eq. (1) and (2) lies with two aspects. First, in Eq. (1), each $\mathbf{x}_{i,j}$ represents a pixel while in Eq. (2), each $\mathbf{x}^*_{i,j}$ represents a series of bounding boxes centered at a pixel. Second, in Eq. (1), each $\mathbf{y}_{i,j}$ represents the category of a pixel. For example, to the task of boundary detection, $\mathbf{y}_{i,j}$ represents the probability of whether a pixel is on the boundary or not. Comparatively, in Eq. (2), each $\mathbf{y}^*_{i,j}$ represents the category of a bounding box. As a consequence, in our network, the overall structural information will be integrated into the network features. As will be seen from experiments on both synthetic and real-world benchmark datasets, the features enhanced by the DDS module are more instance-discriminative than the classic multi-task models.

2.3. Panoptic segmentation

As our proposed algorithm can provide the semantic segmentation map and the instance boundaries, it can also be categorized as a panoptic segmentation (Kirillov et al., 2020) algorithm. Compared with semantic segmentation and instance segmentation, which concentrate

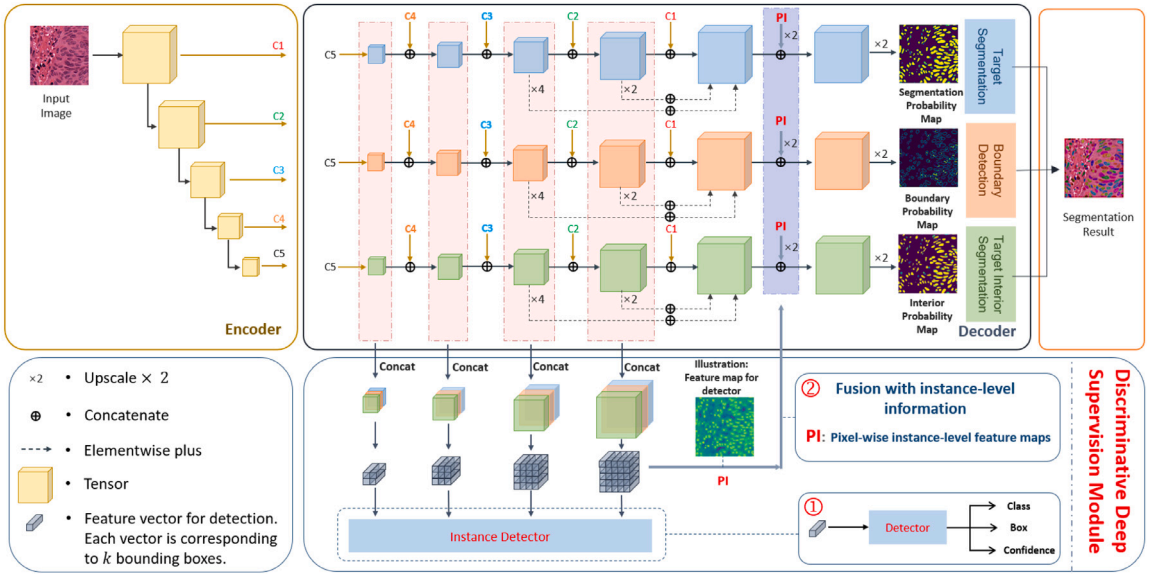


Fig. 3. The network structure of the proposed algorithm. In our network, detection is integrated into a proposal-free instance segmentation network in a deep supervision scheme. The discriminative deep supervision is provided in two folds. In the ① part, a proposal-based detector takes the feature maps generated by the three sub-tasks and injects structural information by passing the gradient through the network. In the ② part, the pixel-wise instance-level feature maps used for detection are passed back to the backbone of the network and merged with the features of the three sub-tasks to help segment the targets directly. Since in the feature maps generated by the detection task, each feature vector corresponds to k (the pixel-wise anchor number) bounding boxes, concatenating these feature maps with the segmentation feature maps provides pixel-wise instance-level information for the proposal-free segmentation backbone network.

on segmenting amorphous image regions and countable objects, panoptic segmentation aims to simultaneously segment countable objects (thing) and uncountable image regions (stuff). Technically, for a pixel in a given image, a panoptic segmentation algorithm predicts its class label if it belongs to stuff while predicting instance class and instance ID if it belongs to thing. The mainstream solutions for the problem can be roughly categorized into two categories, i.e., anchor-free (Yang et al., 2019; Chen et al., 2020), anchor-based algorithms (Xiong et al., 2019; Kirillov et al., 2019). The anchor-free algorithms tend to solve the panoptic segmentation problem in a semantic segmentation-like manner. They detect instances by locating the instance boundary, instance center, instance key points, etc. Anchor-based algorithms tend to locate instances with the help of detection operations. Things and stuff are usually detected separately in these algorithms and then merged. Although remarkable improvement has been made by the mentioned algorithms, segmenting crowded, irregular targets with vague boundaries is still an open problem to the methods in this field (Pang et al., 2019; Chu et al., 2020; Wang et al., 2020b). In this paper, by taking advantage of both methods, we are trying to provide more robust performance against the mentioned problems.

2.4. Deep supervision

The early deep supervision methods (Szegedy et al., 2014; Lee et al., 2015) were proposed to train deeper networks for classification tasks. In these networks, the auxiliary classifiers take the feature maps extracted from the intermediate layers of the network for classification. In this way, the lower layers are trained to be more discriminative to the final target, and the gradients are more smoothly passed through the network. After that, inspired by adding much information to the intermediate layers of the network, segmentation research introduces tasks like boundary detection and centroid detection (Chen et al., 2016; Zhou et al., 2018, 2019; Wang et al., 2020a; He et al., 2021) for performance enhancement. Since deep supervision in this fashion introduces much multi-scale information from complementary information sources, it helps segment targets with better robustness. Although more abundant information is provided, the current manners of deep supervision in segmentation keep using highly correlated low-level auxiliary tasks (like boundary detection, interior detection, and centroid detection) for performance enhancement. However, the accuracy of all these tasks could be easily influenced by the contrast of the target images. To make our proposed algorithm to be more robust against low-contrast images, we introduce our discriminative deep supervision module and adopt a high-level proposal-based detection task (He et al., 2017; Wang et al., 2021) as a new kind of deep supervision. Specifically, to conduct proposal-based detection, three high-level sub-tasks, including bounding box classification, bounding box size adjustment, and Intersection-Over-Union (IOU) estimation, are implemented. As the auxiliary tasks become more complicated and focus more on detecting the whole target, *not only* the textural *but also* the instance structural information is fully exploited. As a result, the features enhanced by our discriminative deep supervision module tend to be more instance-discriminative and robust to low-contrast images. Fig. 2 illustrates the workflow of the proposed algorithm.

3. Method

The intuitive idea of our method is to integrate robust bounding box-centered instance-level information into the proposal-free segmentation algorithm. The overall pipeline and the network architecture of the proposed method are illustrated in Fig. 2 and Fig. 3, respectively. Our DDS-FCN consists of an encoder, a decoder, and the proposed discriminative deep supervision module (DDS module). Of these three components, the encoder and decoder form the backbone of our network. As shown in Fig. 2, the backbone network simultaneously generates semantic segmentation maps, interior segmentation maps, and instance boundary maps. Given an input image, the output probability maps of the sub-networks will be collected by a watershed algorithm for the final instance segmentation. In addition to the encoder and the decoder, the DDS module is further introduced to the network to provide important instance-level information complementary to the semantic segmentation backbone. Concretely, as can be seen in Fig. 3, it first collects the feature maps from all three branches in the decoder and then processes these maps, for instance, detection. At the same time, the learned feature maps are fed back to the decoder to help

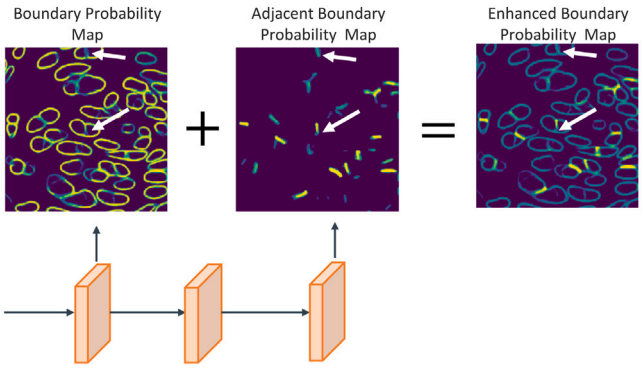


Fig. 4. Boundary detection procedure of our designed network. We found that not all boundaries are with equal difficulty for detection through observation. As can be seen in the first sub-figure (Boundary Probability Map), when all instance boundaries are treated equally and are predicted simultaneously, those uncertain or false predictions usually lie around adjacent areas when cells touch each other (as indicated by the white arrows). After the instance boundary detection, we further use its feature maps to detect only the adjacent boundaries to solve this problem. The final instance boundaries are generated by summing up both predictions to improve the performance on the adjacent areas.

improve the corresponding segmentation performance. In the following subsections, we detail the settings of both the segmentation backbone and the DDS module.

3.1. Segmentation backbone

The segmentation backbone consists of the encoder and the decoder (see Fig. 3). The encoder follows the design of the popular ResNet-101 (He et al., 2016). For the decoder, three branches (corresponding to three sub-tasks) share an identical network architecture (Ronneberger et al., 2015; Lin et al., 2017) but with individual kernel weights. Efficient and straightforward skip connections directly connect the encoder and decoder streams in each branch. The feature maps passed from the encoder are recorded as C1–C5 in Fig. 3. Specially, the incoming channels from the encoder are first concatenated with those from the decoder. Then, as a whole, these channels are passed through a convolutional block (two 3×3 convolutions with batch normalization and ReLU activation). Moreover, to encourage a smooth information flow and better utilization of multi-scale information learned by the network, a deep supervision strategy that combines the intermediate outputs with the final output is adopted (Lee et al., 2015) (see the gray dashed arrows in the decoder of Fig. 3). In our settings, the targets of the semantic segmentation branch are normal semantic segmentation maps of the input image. The target masks of the interior segmentation branch are constructed by eroding and combining the original instance masks. These two branches correspond to the target segmentation sub-network and target interior segmentation sub-network in Fig. 3, respectively.

3.2. Instance boundary detection component

In the three sub-tasks within the segmentation framework, the instance boundary detection, which corresponds to the boundary detection sub-network in Fig. 3, is relatively more complex and essential. Consequently, to improve its performance, we come up with a special design (see Fig. 4). In this design, both the original instance contours and the adjacent boundaries between touching instances are learned by the corresponding sub-network sequentially (see Fig. 4). This setting is based on the observation that the adjacent boundaries between touching instances are usually poorly learned. However, since the instance masks of our method are generated by the watershed algorithm based on the segmentation results and the detected boundaries, the touching boundaries are critical to separate adjacent instances. Therefore, to improve the segmentation accuracy of these boundaries, we guide the network to learn the whole instance boundaries of targets first, and then with the intermediate feature maps, we further force the following networks to infer the location of the touching boundaries. The final instance boundaries are generated by combining the overall instance boundaries and the adjacent boundaries. In our experiments, the adjacent instance boundaries are generated by searching the interaction areas of dilated instance masks.

3.3. Discriminative deep supervision module

To provide instance-discriminative information to the proposal-free segmentation method, we integrate a region proposal network (RPN) (Ren et al., 2015) liked architecture, termed discriminative deep supervision (DDS) module, to the decoder. It is named the DDS module because (1) it is integrated in a deep supervision manner, (2) this module introduces robust structural information to the network, making the learned features of the network more discriminative against vague boundaries.

3.3.1. Overall introduction of the DDS module

Generally, as can be seen in Fig. 3, the DDS module takes the concatenation of the intermediate feature maps of the three sub-tasks in the backbone network as inputs. Then, it puts the processed feature maps into the instance detector and uses it to conduct bounding box classification (Class), bounding box regression (Box), and IoU estimation (Confidence). Specifically, the concatenated feature maps of each sub-task are compressed with a 256-channel convolutional block and 1×1 kernels. In this circumstance, each gray vector in the illustration of the DDS module corresponds to k anchor boxes. (please check (Ren et al., 2015) for more detailed information about the definition of anchor boxes). For the bounding box classification task, a softmax-based classifier is adopted to predict the content in the corresponding anchor box. The bounding box regression task predicts the scale and location difference between a predefined anchor box and the target anchor box. The IoU regression task estimates the interaction ratio between an anchor box between the nearest target anchor box. Besides taking the feature maps, for instance detection, the feature maps generated by the fourth stage of the decoder are further up-scaled and put back to the three branches of the network.

3.3.2. Detail information about the DDS module

Although the structure of the DDS module is to some extent similar to the RPN network, there are five apparent differences. **First**, the DDS module has no Non-Maximum Suppression (NMS) and ROI pooling operation (Ren et al., 2015). Since the integrated detector is only proposed to help inject structural information into the network, it is utilized by training the network with instance-level loss function (anchor box classification, anchor box regression, and anchor box confidence estimation) and passing the pixel-wise instance-level feature maps to the backbone network. As it is not expected to conduct instance detection, there are no NMS operation and ROI pooling operations in the DDS module. This setting naturally alleviates the problem of hard to detect crowded targets and the detection drifting of the detection algorithms. **Second**, since no bounding box refinement is conducted (like what is done in the Mask R-CNN (He et al., 2017)), a hard sample mining strategy that pays more attention to the in-confidently classified negative bounding boxes is adopted in our detector to make our network focus more on the hard samples for better detection accuracy. **Third**, to provide more comprehensive and detailed information of each anchor box, an extra task that estimates the IoU of each box to its best-fitted target instance is introduced. **Fourth**, instead of estimating whether the anchor boxes contain an object or not, the DDS module estimates the categories of each anchor box. **Finally**, we concatenate feature maps from different tasks for the detection module and form a more comprehensive feature space for object detection.

3.4. Loss functions

In the backbone network, for the semantic segmentation, instance segmentation and boundary detection tasks, the same hybrid loss that combines focal loss (Lin et al., 2018) with generalized Dice loss (Sudre et al., 2017) is adopted: $\mathcal{L}_{\mathrm{H}} = \mathcal{L}_{\mathrm{F}} + \mathcal{L}_{\mathrm{GD}}$. Here, $\mathcal{L}_{\mathrm{F}}$ and $\mathcal{L}_{\mathrm{GD}}$ are focal loss and generalized Dice loss, respectively. The focal loss function is

$$\mathcal{L}_{\mathrm{F}} = -\sum_{k=1}^{c}\sum_{i,j}^{m,n} \alpha_k S_{i,j}^k \left(1-p_{i,j}^k\right)^{\gamma} \log p_{i,j}^k, \tag{3}$$

where c is the number of instance classes, m and n corresponds to the sizes of the input images, α_k is the weight balance parameter across different classes, and γ is the focusing parameter used to determine how much an easy pixel is down-weighted. $S_{i,j}^k$ is the one-hot label of pixel (i, j) on the ground truth segmentation maps ($S \in \mathcal{R}^{m\times n\times c}$). It equals to 1 if the pixel belongs to the kth class and 0 otherwise. $p_{i,j}^k$ is the predicted probability of pixel (i, j) coming from the kth class. Denoting the predicted semantic segmentation maps as $S' \in \mathcal{R}^{m\times n\times c}$, the definition of the generalized Dice loss is

$$\mathcal{L}_{\mathrm{GD}} = 1 - 2\cdot \frac{\sum_{k=1}^{c} w_k \sum_{i,j}^{m,n} S_{i,j}^{k\,\prime} S_{i,j}^k}{\sum_{k=1}^{c} w_k \sum_{i,j}^{m,n} \left(S_{i,j}^{k\,\prime} + S_{i,j}^k\right)}, \tag{4}$$

where w_k is the balance term to correct the contribution of each label (Sudre et al., 2017). In this paper, we follow (Sudre et al., 2017) and set $w_k = 1/\left(\sum_{i,j}^{m,n} S_{i,j}^k\right)^2$.

In the DDS module, for bounding box classification and regression, smooth-ℓ_1 loss (Girshick, 2015) is adopted for both box offsets regression ($\mathcal{L}_{\mathrm{BR}}$) and IOU regression ($\mathcal{L}_{\mathrm{IOU}}$). A multi-class focal loss is adopted for bounding box classification ($\mathcal{L}_{\mathrm{BC}}$).

By summing up the losses of all the sub-tasks, we finally obtain our loss function for the whole network

$$\begin{aligned}\mathcal{L} = {} & \mathcal{L}_{\mathrm{SEG}} + \beta_1\mathcal{L}_{\mathrm{BD}} + \beta_2\mathcal{L}_{\text{I-SEG}} \\ & + \beta_3\mathcal{L}_{\mathrm{BR}} + \beta_4\mathcal{L}_{\mathrm{IOU}} + \beta_5\mathcal{L}_{\mathrm{BC}},\end{aligned} \tag{5}$$

where $\mathcal{L}_{\mathrm{SEG}}$, $\mathcal{L}_{\mathrm{BD}}$, and $\mathcal{L}_{\text{I-SEG}}$ are the semantic segmentation loss, boundary detection loss, and instance interior segmentation loss, respectively. $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are parameters used to balance the importance of each term. Through experiments and observations, we have two guidelines for setting these parameters. (1) Setting the weights to make the losses on the backbone network roughly equal to the sum of the detection loss creates a balance between segmentation and detection. (2) Setting the sizes of bounding boxes according to the size of the targets makes the training stable.

3.5. Implementation details

3.5.1. Label generation

With the given instance-level sample masks, the instance boundary, instance adjacent boundary, the instance interior mask, and the corresponding bounding boxes can be generated accordingly without extra human labeling.

To generate the instance contours, we first take the mask of a specific instance and find the contour of the instance. Secondly, by conducting binary dilation twice, we generate the instance contour for that instance. Thirdly, we add the contours of all instances and merge the sub-instance maps into a unit figure. Finally, we set the non-zero components as positive and generate the label for the instance contours. To generate the label of the adjacent boundaries, we take the unit figure generated in the third step of the instance contour label generation procedure. Then, we keep the components that exist overlap as positive. Finally, we conduct binary dilation once and obtain the label map. For the instance interior mask generation, we simply conduct a simple image deduction between the instance mask and the instance boundary. Finally, for the instance bounding boxes, we take the corresponding code in Mask-RCNN[2] and adopt the instance segmentation masks as input for their generation.

3.5.2. Model setting

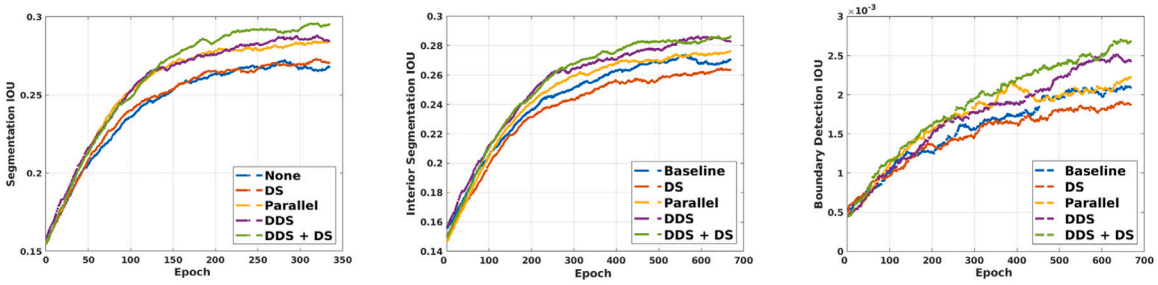
We use the PyTorch platform for implementations. For training, the patch size is set to $512 \times 512 \times 3$ for the nuclei segmentation and $256 \times 256 \times 3$ for pelvic CT image segmentation. Except for U-Net (Ronneberger et al., 2015), all the backbones of the compared networks are initialized with well-trained kernel weights. Specifically, the weights of Multi-Task U-Net, DeepLabV3+, PSPNett are pre-trained on ImageNet (http://www.image-net.org) while the weights of Mask RCNN is pre-trained on MSCOCO (Lin et al., 2014), etc. The Adam optimizer with default parameters is used to train the network. The batch sizes for nuclei segmentation and pelvic organ segmentation are 4 and 20, respectively. For the former dataset, we first fix the backbone of the network and train it for 200 epochs with a learning rate of 0.001. Next, we decrease the learning rate to 0.0001 and train the whole network for 600 epochs. Finally, we further decreased the learning rate to 0.00001 and trained the network for 1200 epochs. The training strategy of pelvic organ segmentation is the same as nuclei segmentation, except that the training epoch for different stages is 5, 20 and 20, respectively. We randomly divided the two datasets by the ratio of 5 : 2 : 3 for training, validation, and testing. For the hyperparameters, we set $\beta_1 = 1, \beta_2 = 1, \beta_3 = 1.2, \beta_4 = 1.2, \beta_5 = 2.5$ for the nuclei dataset. The anchor box sizes are $8, 16, 32, 64$. For the prostate segmentation dataset, we set $\beta_1 = 0.2, \beta_2 = 0.5, \beta_3 = 1.2, \beta_4 = 3, \beta_5 = 4$. The anchor box sizes are $16, 32, 56, 80, 110$. Through experiments and observations, we have two guidelines for setting these parameters. (1) Setting the weights to make the losses on the backbone network roughly equal to the sum of the detection loss creates a balance between segmentation and detection. (2) Setting the sizes of bounding boxes according to the size of the targets makes the training stable.

For data augmentation, heavy data augmentation (including random zooming, rotation, flipping, channel shifting, elastic transform, and adding noise) is employed in the nuclei dataset to avoid overfitting. For the pelvic CT image dataset, since the location of the target organs is relatively fixed, light augmentation is used (including random zooming, light rotation, up–down flipping, and noise adding). The augmentation parameters are set as follow: for random zooming, the minimal scale of zooming is 0.5 and the maximal scale is 2. For random rotation, each image would be rotated for $0, 90, 180$, and 270 degree. For flipping, each input have 50% probability of conducting left–right and up–down flipping. For channel shifting, each image has a probability of 50% to conduct random channel shifting. For elastic transform, each image has a probability of 25% of doing this augmentation. The parameter α and δ for Gaussian filter generation is 2000 and 30, respectively. For the adding noise operation, speckle noise is randomly added with 50% probability.

4. Experiments

In this section, we evaluate the performance of our proposed algorithm, especially the effectiveness of the discriminative deep supervision module on two real-world datasets, i.e., a cell nucleus digital microscopic image dataset and a pelvic organ CT image dataset. We further construct a toy dataset to validate the effectiveness of the DDS module in extracting structural information of instances and the robustness of our proposed model against blurry boundaries.

[2] https://github.com/matterport/Mask_RCNN.



(a) Target Segmentation (b) Interior Segmentation (c) Adjacent Boundary Detection

Fig. 5. Validation curves of the compared algorithms. Sub-figure (a–c) are the validation curves of the target segmentation branch, interior segmentation branch and the adjacent boundary detection branch, respectively. Six algorithms are compared. Among them, Baseline indicates the network with only the backbone architecture, DS indicates the classic deep supervision enhanced baseline, Parallel indicates adding detection as a parallel task to the backbone network, DDS indicates adding detection in a deep supervision manner to the network, DDS+DS indicates further adding classic deep supervision to the DDS network.

4.1. Evaluation datasets

Nuclear segmentation dataset. The nuclear segmentation dataset is a digital microscopic tissue image challenge dataset for nuclear morphometrics and other analysis in computational pathology (Kumar et al., 2020). The dataset was obtained by carefully annotating tissue images of several patients with tumors of different organs (i.e., prostate, breast, colon, liver, bladder, kidney, and stomach). It is a binary segmentation dataset containing around 22,000 nuclear boundary annotations in the learning set. Both benign and diseased tissue samples are included to ensure nuclear appearance diversity. The training images are collected from different hospitals, improving the appearance variation of the collected images.More information about the MoNuSeg challenge and the dataset can be found in the challenge website.[3]

Pelvic CT image segmentation dataset. The pelvic CT image dataset is a multi-class segmentation dataset, and it is acquired by the North Carolina Cancer Hospital, which includes 339 CT scans from prostate cancer patients. For this dataset, three organs (i.e., prostate, bladder, and rectum) are being segmented in each CT image. An experienced radiation oncologist manually delineates the target organs to serve as the ground truth in our training and testing procedure. This dataset is included to enrich the testing scenario (CT images). It is also selected because the target size in different slices of prostate CT image segmentation various more intensive than that in the nuclear images. To segment the 3D pelvic CT images, the adjacent 3 slices are combined as the input of our network. The segmentation results of the consecutive slices are concatenated to form the final out put of a subject.

Synthetic dataset. We further design a synthetic dataset to illustrate the structure information extracting capability of the DDS module. The main idea of the experiment is to create foreground targets with similar but distinguishable plain textures so that classic algorithms can tell apart targets with the help of texture information in this circumstance. Then, we add noise to the dataset to blur the boundary and hard to tell apart. In this circumstance, only structure information can help to detect the foreground target. Finally, we test and compare the performance of our proposed algorithm with and without the DDS module to evaluate its effectiveness in exploiting the structure information of the detection targets.

For the construction of the dataset, the target instances include three typical shapes, i.e., square, circle, and triangle. In each of the generated 128×128 images, we include three objects with random shapes and with sizes around 28 pixels. The color within each object is consistent. Moreover, the colors for different objects are different, but each RGB channel's difference is restricted to the range of $[10, 15]$. With this setting, we construct instances with similar textures. The images for training, validation, and testing are 8000, 1000, and 2000, respectively. We train and test the network two times, and neither data augmentation is included in both processes. The networks are trained regularly without special operations for the first round of training and testing. However, we add Gaussian noise with a variance of 0.008 to the input images for the second round. Also, we smooth the images with a Gaussian kernel with a kernel size of 8. By doing this, the boundaries at the touching regions become vague and harder to distinguish.

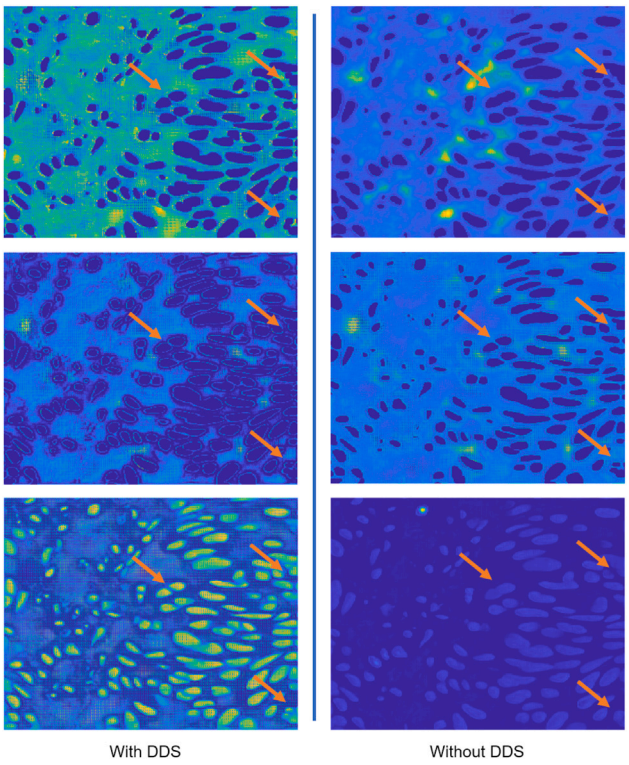
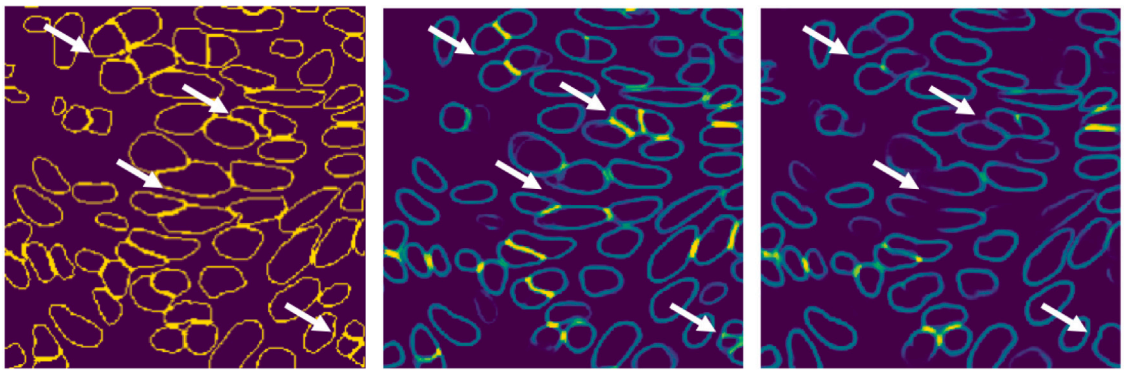


Fig. 6. Feature discriminative capacity comparison. We visualize representative feature maps at the second-to-last convolutional layer of the interior segmentation branch learned by the network with and without enhancing the DDS module. Through the figure, we can easily find that the DDS module does help improve the discriminative capability of the network.

4.2. Ablation experiments

We first showcase the effectiveness of the proposed method, especially the DDS module, with an extensive quantitative and qualitative comparison of the nuclei dataset.

[3] https://monuseg.grand-challenge.org



(a) Ground Truth (b) With DDS (c) Without DDS

Fig. 7. Visualization of the learned instance boundaries of the proposed networks with and without DDS module. As pointed out by the white arrows, the ambiguous boundaries around the adjacent areas are better detected with the DDS module.

4.2.1. Quantitative evaluation

This part evaluates the necessity of the two main components: (a) the discriminative deep supervision module and (b) the pixel-wise instance-level information introduction mechanism. Besides, we also evaluate the effectiveness of the classic deep supervision and the design of using detection as a parallel task in the instance segmentation scenario. To conduct the evaluation, we construct six networks and show the results of different combinations of these components. Four criteria, i.e., the F1 scores (F1), instance Dice ratios (Dice), Hausdorff distances (HD) and Aggregated Jaccard Index (AJI) are reported in Table 1. The results indicate that (1) detection provides important complementary information to the backbone FCN for instance segmentation. As shown in Table 1, even integrated in a parallel fashion, the detection task improves the F1 score significantly for almost 3%; (2) comparing with the baseline network, the improvement provided by classic deep supervision is limited; (3) integrating the detection task in a deep supervision manner and connects different sub-tasks comprehensively is more beneficial than using them as separate parallel sub-tasks. The improvement introduced by this design is 2.2% on F1 score, 1.2% on Dice, and 0.85 mm on HD; (4) combining the classic deep supervision, DDS module, and IP enhancement provides the best performance. The network in this form improves the baseline for more than 5%, 3%, and 2.6 mm on F1 score, Dice, and HD, respectively. Also, it is worth noting that the parameter number of the baseline algorithm is 191.2 million, while the parameter number of the proposed algorithm is only 195.4 million. The DDS module brings significant improvement with only limited parameter number increase.

In this paper, to save training and testing time, random sample division instead of k-fold cross validation is adopted for algorithm performance evaluation. It is worth knowing that dividing the datasets into training and testing sets is also quite popular in the field of nuclei and prostate segmentation scenarios. Many papers with high citations, like DCAN, Z-Net, etc, (Chen et al., 2016; Schmidt et al., 2018) have adopted this kind of setting. Moreover, to prove that the results tested with the current fashion are consistent with the results of k-fold cross validation, we conduct 5-fold cross-validation and compare the results of two state-of-the-art algorithms, i.e., Multi-task U-Net and Mask-RCNN. Specifically, in each round of testing, one of the five folds of samples are used for testing and the remaining samples are used for training and validation (with a ratio of 5:1). We repeat this procedure five times to ensure that all the five folds of samples are traversed as the testing set. The F1 score of the two methods are 84.27.4, 86.55.2 while the F1 score 90.83.8. As we can see, (1) the results again prove the effectiveness of the proposed algorithm; (2) the proposed modules tend to improve the stability of the corresponding algorithm.

Table 1
Ablation study of the DDS-FCN. Here, BL denotes the baseline network. In the experiment, it is constructed by only the encoder and the decoder as shown in Fig. 3. DS denotes classic deep supervision, DP denotes detection as a parallel task, DDS denotes detection as deep supervision, PI denotes pixel-wise instance-level information.

Method	F1 (%)	Dice (%)	HD	AJI (%)
BL	87.0	83.0	8.11	60.97
BL+DS	87.1	83.1	7.54	60.09
BL+DP+DS	90.0	84.2	6.61	64.56
BL+DDS	91.7	85.4	5.95	67.01
BL+DDS+DS	92.2	85.4	5.76	67.47
BL+DDS+DS+PI	**92.5**	**86.1**	**5.51**	**68.20**

4.2.2. Learning curve comparison

To dynamically evaluate the effect of the DDS module on the training process in detail, we report the smoothed training and validation IOU values of target interior segmentation and adjacent boundary detection sub-tasks in Fig. 5. From the figures, we notice that the green and purple curves, which represent the networks with DDS modules, maintain better performance than the other competitors in most cases in both the training and the validation sets. Moreover, the combination of the classic deep supervision and the proposed DDS module provides an even more consistent improvement. The results in this subsection verify that the integrated DDS module does boost the performance of the backbone network.

4.2.3. Discriminative capability comparison on the learned features

To compare the discriminative capability of the features learned by networks with and without the DDS module, we visualize the feature maps at the second-to-last convolution layer of the interior segmentation branch. As we can see in Fig. 6, in the learned feature maps, (1) the touching samples with blurry boundaries (pointed out by the upper two arrows), (2) the crowded and small samples (the lower arrow on the right corner) are better recognized by the feature maps enhanced by the DDS module, indicating better discriminative capability.

4.2.4. Boundary visualization

We further illustrate the learned boundaries of the baseline networks with and without the DDS module to assess the effect of the module. As we can see in Fig. 7, the boundaries that are clean and clear can be easily and accurately detected by both networks. However, the vague ones, which typically lie around the adjacent areas between nuclei, are less discriminative and harder to be detected than the other counterparts. Under this circumstance, the network with the DDS module provides a more robust and more accurate response on those areas, thus providing better guidance to the network for the corresponding nuclei.

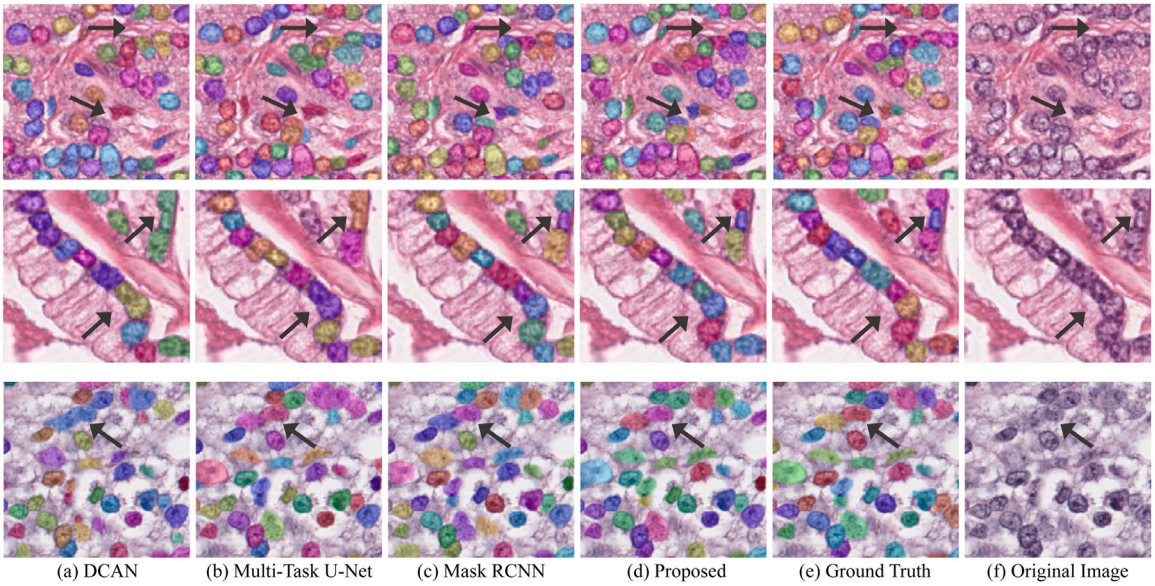


Fig. 8. Visualization of segmentation results of four compared algorithms on three representative samples with complex and ambiguous boundaries.

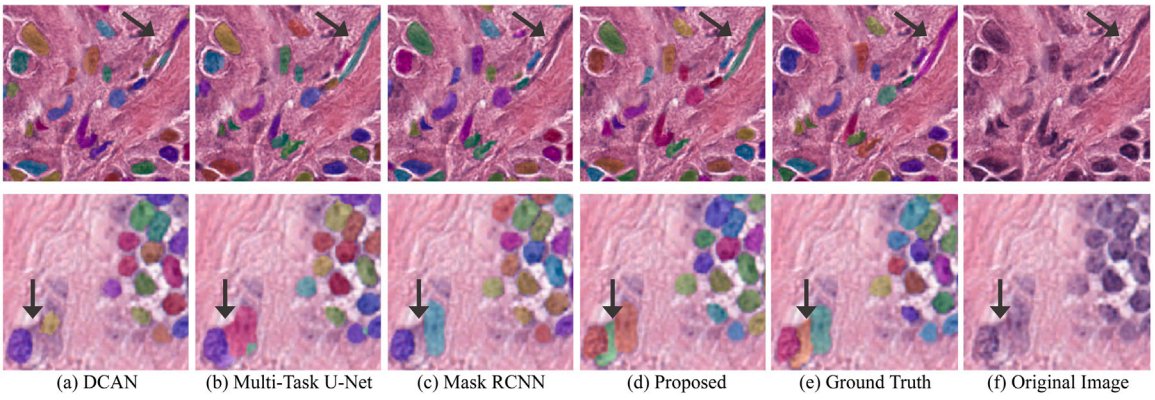


Fig. 9. Visualization of segmentation results of four compared algorithms on two representative samples with irregular shapes and sizes.

Table 2
Comparison with the state-of-the-art methods on the nuclei segmentation datasets.

Method	F1 (%)	Dice (%)	HD	AJI (%)
DCAN	85.8	80.3	8.86	60.00
Multi-Task U-Net	88.1	83.9	7.14	62.52
Mask-RCNN	89.8	85.1	6.21	65.49
Proposed	**92.5**	**86.1**	**5.51**	**68.20**

4.3. Comparison with the state-of-the-art methods

This part compares the proposed algorithm with the state-of-the-art methods on two datasets, i.e., the nuclei dataset and the pelvic CT image dataset.

4.3.1. The nuclei dataset

To the first dataset, three champion methods, i.e., DCAN (Chen et al., 2016) (the top method of Gland Segmentation Challenge Contest[4] and Segmentation of Nuclei in Digital Pathology Images), Multi-task U-Net[5] (the top method for Kaggle 2018 Data Science Bowl challenge,[6]) and Mask R-CNN[7] (MSCOCO 2016 challenge winners (Lin et al., 2014)) are compared. Among these methods, the first two represent the proposal-free methods. Although simple and easy to understand, through careful pre- and post- processing, algorithms in this branch have won the competition of many nuclei segmentation challenges (including the MoNuSeg 2018 challenge (Kumar et al., 2020)). Comparatively, Mask R-CNN is selected as the representative for the proposal-based method. Since we focus on network performance comparison, we follow the winning parameter settings of the compared algorithms and share the same pre- and post- processing. Specifically, for the multi-task U-Net, we ensemble Res-101 (He et al., 2016) and Dense-121 (Huang et al., 2017) by averaging the pixel-wise prediction of the two models on each branch for results estimation. For Mask R-CNN, the parameters are set according to the suggestion of the 3rd place winner of the Kaggle Bowel competition.[8]

In Table 2, one can see that our algorithm outperforms the state-of-the-art algorithms in all three metrics. Concretely, our proposed algorithm surpassed the second-best algorithm for 2.7% on F1 score,

[4] https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest

[5] https://github.com/selimsef/dsb2018_topcoders

[6] https://www.kaggle.com/c/data-science-bowl-2018

[7] https://github.com/matterport/Mask_RCNN

[8] https://www.kaggle.com/c/data-science-bowl-2018/discussion/56393

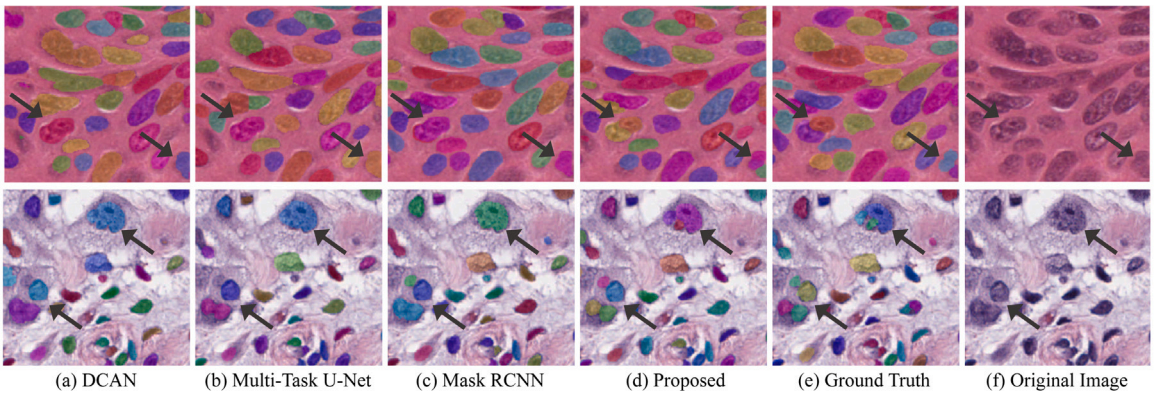


Fig. 10. Segmentation results visualization of the compared algorithms on two representative samples with the tiny targets densely distributed. In these figures, the randomly colored ovals indicate predicted nuclei masks. The segmentation results of (a) DCAN, (b) Multi-task U-Net, (c) Mask R-CNN, and (d) the proposed method, as well as (e) the ground truth segmentation, are illustrated. The original input images are illustrated in sub-figure (f). The dark arrows indicate the area where representative nuclei locate.

Table 3
Comparison with the state-of-the-art methods on the pelvic CT image dataset.

Method	Prostate	Bladder	Rectum
	DSC (%)		
U-Net	85.1	90.7	81.8
DCAN	86.3	91.8	82.2
DeepLabV3+	86.8	92.5	83.3
PSPNet	87.1	93.1	83.3
Proposed	**88.1**	**93.9**	**84.7**
Method	Prostate	Bladder	Rectum
	ASD (mm)		
U-Net	1.74	1.55	1.88
DCAN	1.56	1.32	1.79
DeepLabV3+	1.54	1.27	1.68
PSPNet	1.45	1.26	1.73
Proposed	**1.35**	**1.16**	**1.49**

Table 4
Performance comparison of networks with and without the DDS module on synthetic datasets with and without intense Gaussian noise. The instance boundaries are stable and clear in the dataset without Gaussian noise. Comparatively, the instance boundaries are ambiguous in images with Gaussian noise. Average precision (AP) of the compared networks is reported.

	No Gaussian noise	With Gaussian noise
With DDS	**90.0%**	**71.3%**
Without DDS	89.6%	61.7%

1% on Dice ratio, 11.2% on Hausdorff distances, and 2.71% on AJI, respectively.

Moreover, to better showcase the advantages of the proposed algorithm, in Figs. 10, 8 and 9, we illustrate the segmentation results in three kinds of representative circumstances. Specifically, Fig. 10 corresponds to the circumstances with tiny crowded targets. Fig. 8 corresponds to the circumstances with blurry and complex boundaries. Fig. 9 corresponds to the circumstances with irregular target shapes and sizes. The results suggest that our algorithm outperforms the Mask R-CNN on segmenting tiny and densely-distributed nuclei with irregular shapes and surpasses the proposal-free methods for segmenting touching nuclei with vague and ambiguous boundaries.

4.3.2. The pelvic CT image dataset

Four state-of-the-art semantic segmentation algorithms, including U-Net (Ronneberger et al., 2015), DCAN (Chen et al., 2016), DeepLabV3+ (Chen et al., 2018), and PSPNet (Zhao et al., 2017) are used for comparison. From Table 3, we observe that, although all the compared state-of-the-art algorithms perform reasonably well on this dataset, our proposed method still outperforms those compared algorithms for more than 1% on Dice ratio and 7% on average surface distance in almost all the three target organs. One reason for the better results may be the fact that our algorithm can extract organ-level structural information, thus is more robust to the blurry and ambiguous boundaries.

In Fig. 11, we further illustrate the false predictions of the compared algorithms to help analyze the mechanism of effectiveness. Specifically, in columns (a, b, d, e), we record the false predictions of the compared algorithms by calculating the difference between the estimated segmentation and the ground truth of the target organs. The less the false prediction is, the better performance it corresponds to. As we can see from the figure, the pelvic CT image dataset has comparatively lower contrast than the nuclei images. The false predictions mainly lie around the boundaries of target organs. With the help of the structural information collection capability, our proposed algorithm provides better boundary localization performance, especially on the blurry boundaries between target organs.

4.4. Effectiveness of the pixel-wise instance-level information

To justify the effectiveness of the pixel-wise instance-level information and intuitively reveal why the DDS module is vital for improving the performance of the proposal-free instance segmentation methods, we further design two toy binary instance segmentation datasets with and without blurry and ambiguous boundaries. By conducting experiments on these datasets, we explore the properties of the DDS module, thus revealing its working mechanism.

In Table 4, we report the average precision (AP) value of the networks on the datasets with and without Gaussian noise, respectively. From the table, we can see that, when the images are clear, the performance of the two networks is satisfactory and comparable. However, after adding the Gaussian noise and making the boundaries between instances ambiguous, although the performances of both networks decrease drastically, the one with the DDS module performs 10% better than its counterpart in terms of AP. To explain this phenomenon, we illustrate the learned boundary maps of the network without the DDS module and a representative detection feature map from the network with the DDS module in Fig. 12. The figure shows that, compared with the instance boundaries, detection feature maps are more robust indicators for instances in low-contrast images with blurry boundaries.

5. Discussion

As can be seen from the results of experiments on the three compared datasets, the proposed algorithm has integrated both the advantages of proposal-free and proposal-based methods. As a result,

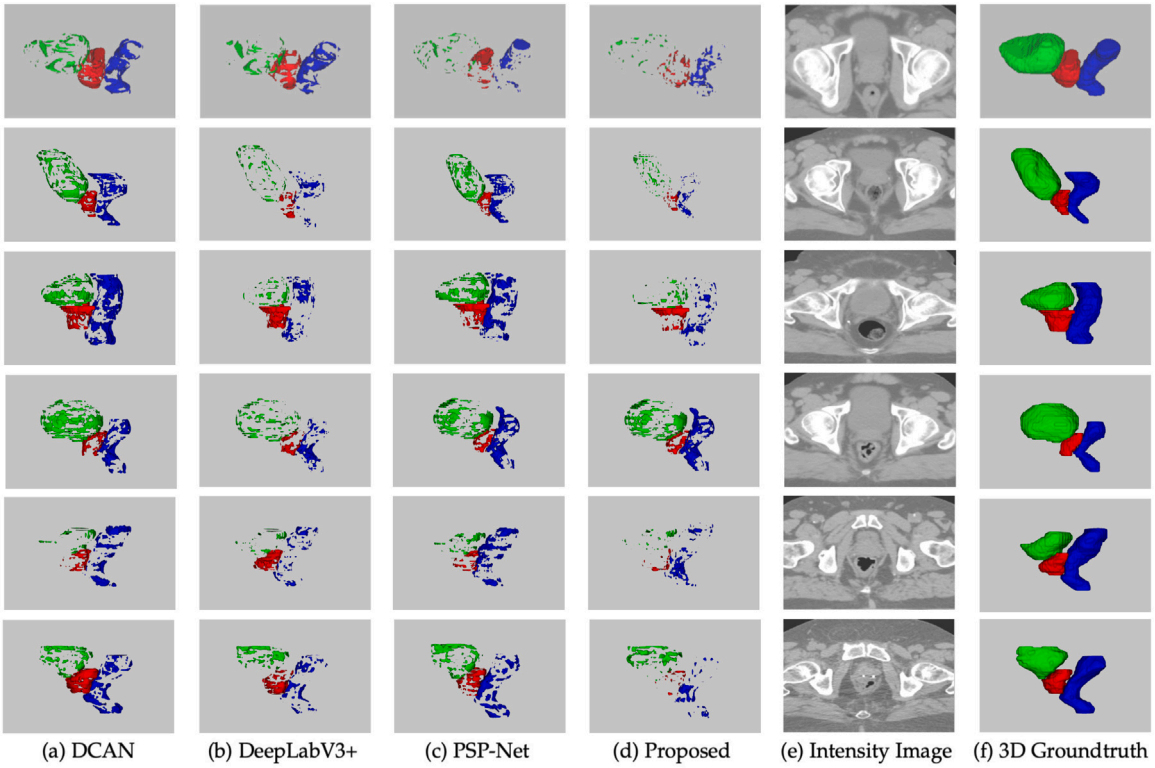


Fig. 11. Results illustration on the pelvic CT image dataset. The 3D difference map, intensity image, and the 3D ground truth of representative samples are reported. From the first to the fourth column are the error maps of the compared algorithms. In these maps, green, red, and blue fractions indicate false estimations on bladder, prostate, and rectum, respectively. The cleaner an image is, the better the performance will be. Images in the fifth column are representative CT intensity image maps in the axial direction. The last column images are the ground truth 3D segmentation for the representative samples.

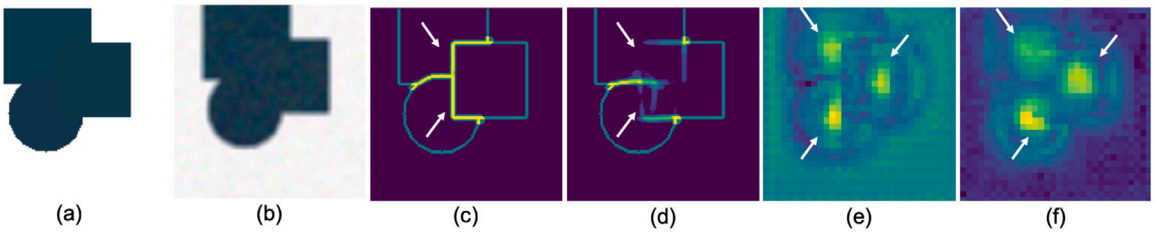


Fig. 12. Robustness comparison between the boundary maps and the detection feature maps for instance segmentation with (a, c, e) and without ambiguous boundaries (b, d, f). This figure illustrates the intensity image map (a, b), the learned boundary maps (c, d) and the learned detection feature maps (e, f) before (a, c, e) and after (b, d, f) blurring operation. In the experiment, the instance boundaries are clean and clear before the blurring operation and indistinctive afterward. The boundary maps are generated by the baseline network while the detection feature maps are generated by the DDS module enhanced network. As we can see, the detection feature maps are more robust to noise and ambiguous boundaries for instance localization. This illustrates the source of effectiveness of our proposed algorithm.

it performs properly on images with crowded and irregular targets and images with low contrast and blurry boundaries. Moreover, since the proposed DDS module takes the intermediate feature maps of the three branches in the backbone network as inputs, it places little extra parameters and computation burden on the existing network. Also, since all the labels of the additional tasks can be generated automatically according to the instance masks, the newly proposed framework introduces little extra human labor against the existing algorithms. However, it also has its limitations. The main limitation of our algorithm is the complex network architecture. This could inevitably influence the generalization capability, network training difficulty, and computational and memory consumption. Specifically, the three branch architecture introduces a triple number of parameters to the decoder of the network, which enlarges the risk of overfitting and cost non-negligible computational resource consumption. Furthermore, since there are six terms in the loss function of the network, searching for the best hyper-parameter in an extensive range would also increase the difficulty of network training. Moreover, compared with the instance segmentation algorithm like Mask RCNN and DCAN, the proposed algorithm segments target instances in a two-step manner, i.e., generating the semantic segmentation, instance boundary, and instance interior first and then segment the instances with the watershed algorithm. This manner isolates the final segmentation procedure from representation learning, thus limiting the performance of the proposed algorithm. Moreover, the proposal-based detection manner of the DDS module still limits the capability of capturing structure information of irregular targets. The recently proposed anchor-free algorithms are potentially more appropriate for the current framework. In the future, more research would be done to replace the anchor based detection module with an anchor-free detection module to further improve the proposed algorithm's performance.

6. Conclusion

This paper proposed a novel discriminative deep supervision-based fully convolutional network (DDS-FCN) to make semantic instance

segmentation robust against crowded samples, significant shape variations, and low-contrast medical images. To introduce instance-sensitive structural information to our proposal-free segmentation method and eliminate the limitation of the proposal-based segmentation mechanism, we introduced a novel discriminative deep supervision module by integrating a detection task into the segmentation backbone network in a deep supervision manner. The resulting algorithm was both discriminative to instance detection and stable against variations of target sizes. Without the pixel-wise instance-level feature maps and the regular deep supervision operation, the intermediate version of our algorithm had won the second and sixth prizes in the NCI-MICCAI 2018 digital pathology segmentation of nuclei challenge and the Multi-Organ Nucleus Segmentation Challenge.[9] The challenge results, together with the experimental results in the paper, validate the superior performance of the proposed algorithm.

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.

Acknowledgments

J. Yin, S. Zhou, X, Liu and E, Zhu were supported in part by the National Key R&D Program of China 2020AAA0107100, and the National Natural Science Foundation of China under Grant 62006237, 61872371. D. Shen was supported in part by NIH grant (CA206100). E. Adeli was supported in part by NIH Grant (AA026762).

References

Arnab, A., Torr, P.H., 2017. Pixelwise instance segmentation with a dynamically instantiated network. In: CVPR, Vol. 1. (2), p. 5.

Bai, M., Urtasun, R., 2017. Deep watershed transform for instance segmentation. In: CVPR. IEEE, pp. 2858–2866.

Chen, Q., Cheng, A., He, X., Wang, P., Cheng, J., 2020. Spatialflow: Bridging all tasks for panoptic segmentation. IEEE Trans. Circuits Syst. Video Technol. PP (99), 1.

Chen, L.-C., Hermans, A., Papandreou, G., Schroff, F., Wang, P., Adam, H., 2017. Masklab: Instance segmentation by refining object detection with semantic and direction features. arXiv preprint arXiv:1712.04837.

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

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv:1802.02611.

Chu, X., Zheng, A., Zhang, X., Sun, J., 2020. Detection in crowded scenes: One proposal, multiple predictions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12214–12223.

Dai, J., He, K., Li, Y., Ren, S., Sun, J., 2016a. Instance-sensitive fully convolutional networks. In: ECCV. Springer, pp. 534–549.

Dai, J., He, K., Sun, J., 2016b. Instance-aware semantic segmentation via multi-task network cascades. In: CVPR. pp. 3150–3158.

Dai, J., Li, Y., He, K., Sun, J., 2016c. R-fcn: Object detection via region-based fully convolutional networks. In: NIPS. pp. 379–387.

De Brabandere, B., Neven, D., Van Gool, L., 2017. Semantic instance segmentation with a discriminative loss function. arXiv preprint arXiv:1708.02551.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. IJCV 88 (2), 303–338.

Fathi, A., Wojna, Z., Rathod, V., Wang, P., Song, H.O., Guadarrama, S., Murphy, K.P., 2017. Semantic instance segmentation via deep metric learning. arXiv preprint arXiv:1703.10277.

Girshick, R., 2015. Fast r-cnn. In: ICCV. pp. 1440–1448.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: ICCV. IEEE, pp. 2980–2988.

He, K., Lian, C., Zhang, B., Zhang, X., Cao, X., Nie, D., Gao, Y., Zhang, J., Shen, D., 2021. HF-UNet: Learning hierarchically inter-task relevance in multi-task U-net for accurate prostate segmentation in CT images. IEEE Trans. Med. Imaging.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: CVPR. pp. 770–778.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: CVPR, Vol. 1. (2), p. 3.

Huang, J., Shen, Y., Shen, D., Ke, J., 2021. CA 2.5-net nuclei segmentation framework with a microscopy cell benchmark collection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 445–454.

Kendall, A., Gal, Y., Cipolla, R., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR.

Kirillov, A., Girshick, R., He, K., Dollar, P., 2019. Panoptic feature pyramid networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P., 2020. Panoptic segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Kirillov, A., Levinkov, E., Andres, B., Savchynskyy, B., Rother, C., 2017. Instancecut: from edges to instances with multicut. In: CVPR, Vol. 3. p. 9.

Kumar, N., Verma, R., Anand, D., 2020. A multi-organ nucleus segmentation challenge. IEEE Trans. Med. Imaging 39 (5), 1380–1391.

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z., 2015. Deeply-supervised nets. In: Artificial Intelligence and Statistics. pp. 562–570.

Lessmann, N., van Ginneken, B., de Jong, P.A., Išgum, I., 2018. Iterative fully convolutional neural networks for automatic vertebra segmentation. arXiv preprint arXiv:1804.04383.

Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y., 2017. Fully convolutional instance-aware semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 4438–4446.

Lin, T.-Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J., 2017. Feature pyramid networks for object detection. In: CVPR, Vol. 1. (2), p. 4.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2018. Focal loss for dense object detection. IEEE TPAMI.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: ECCV. Springer, pp. 740–755.

Liu, S., Jia, J., Fidler, S., Urtasun, R., 2017. Sgn: Sequential grouping networks for instance segmentation. In: ICCV.

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path aggregation network for instance segmentation. In: CVPR. pp. 8759–8768.

Novotny, D., Albanie, S., Larlus, D., Vedaldi, A., 2018. Semi-convolutional operators for instance segmentation. In: ECCV. Springer, pp. 89–105.

Pang, J., Li, C., Shi, J., Xu, Z., Feng, H., 2019. $\mathcal{R}^2$-CNN: fast tiny object detection in large-scale remote sensing images. IEEE Trans. Geosci. Remote Sens. 57 (8), 5512–5524.

Papandreou, G., Zhu, T., Chen, L.-C., Gidaris, S., Tompson, J., Murphy, K., 2018. PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. arXiv preprint arXiv:1803.08225.

Pinheiro, P.O., Collobert, R., Dollár, P., 2015. Learning to segment object candidates. In: NIPS. pp. 1990–1998.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer, pp. 234–241.

Roth, H.R., Lu, L., Lay, N., Harrison, A.P., Farag, A., Sohn, A., Summers, R.M., 2018. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Med. Image Anal. 45, 94–107.

Salvador, A., Bellver, M., Baradad, M., Marqués, F., Torres, J., Giro-i Nieto, X., 2017. Recurrent neural networks for semantic instance segmentation. arXiv preprint arXiv:1712.00617.

Schmidt, U., Weigert, M., Broaddus, C., Myers, G., 2018. Cell detection with star-convex polygons. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 265–273.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 240–248.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Rabinovich, A., 2014. Going deeper with convolutions. IEEE Comput. Soc..

Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M., 2021. Scaled-yolov4: Scaling cross stage partial network. In: Proceedings of the IEEE/Cvf Conference on Computer Vision and Pattern Recognition. pp. 13029–13038.

Wang, S., He, K., Nie, D., Zhou, S., Gao, Y., Shen, D., 2019. CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation. Med. Image Anal. 54, 168–178.

[9] https://monuseg.grand-challenge.org/Results/

Wang, S., Liu, M., Lian, J., Shen, D., 2020a. Boundary coding representation for organ segmentation in prostate cancer radiotherapy. IEEE Trans. Med. Imaging 40 (1), 310–320.

Wang, Y., Xie, H., Zha, Z.-J., Xing, M., Fu, Z., Zhang, Y., 2020b. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11753–11762.

Xiong, Y., Liao, R., Zhao, H., Hu, R., Urtasun, R., 2019. UPSNet: A unified panoptic segmentation network. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Xu, Y., Li, Y., Wang, Y., Liu, M., Fan, Y., Lai, M., Eric, I., Chang, C., 2017. Gland instance segmentation using deep multichannel neural networks. IEEE Trans. Biomed. Eng. 64 (12), 2901–2912.

Yang, T.-J., Collins, M.D., Zhu, Y., Hwang, J.-J., Liu, T., Zhang, X., Sze, V., Papandreou, G., Chen, L.-C., 2019. DeeperLab: Single-shot image parser. ArXiv Preprint ArXiv:1902.05093.

Zagoruyko, S., Lerer, A., Lin, T.-Y., Pinheiro, P.O., Gross, S., Chintala, S., Dollár, P., 2016. A multipath network for object detection. arXiv preprint arXiv:1604.02135.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: CVPR. pp. 2881–2890.

Zhou, S., Nie, D., Adeli, E., Gao, Y., Wang, L., Yin, J., Shen, D., 2018. Fine-grained segmentation using hierarchical dilated neural networks. In: MICCAI. Springer, pp. 488–496.

Zhou, S., Nie, D., Adeli, E., Yin, J., Lian, J., Shen, D., 2019. High-resolution encoder–decoder networks for low-contrast medical image segmentation. IEEE Trans. Image Process. 29, 461–475.