# Domain Adaptive Box-Supervised Instance Segmentation Network for Mitosis Detection

Yonghui Li, Yao Xue, Liangfu Li, Xingjun Zhang, *Member, IEEE*, and Xueming Qian, *Member, IEEE*

***Abstract*— The number of mitotic cells present in histopathological slides is an important predictor of tumor proliferation in the diagnosis of breast cancer. However, the current approaches can hardly perform precise pixel-level prediction for mitosis datasets with only weak labels (i.e., only provide the centroid location of mitotic cells), and take no account of the large domain gap across histopathological slides from different pathology laboratories. In this work, we propose a Domain adaptive Box-supervised Instance segmentation Network (DBIN) to address the above issues. In DBIN, we propose a high-performance Box-supervised Instance-Aware (BIA) head with the core idea of redesigning three box-supervised mask loss terms. Furthermore, we add a Pseudo-Mask-supervised Semantic (PMS) head for enriching characteristics extracted from underlying feature maps. Besides, we align the pixel-level feature distributions between source and target domains by a Cross-Domain Adaptive Module (CDAM), so as to adapt the detector learned from one lab can work well on unlabeled data from another lab. The proposed method achieves state-of-the-art performance across four mainstream datasets. A series of analysis and experiments show that our proposed BIA and PMS head can accomplish mitosis pixel-wise localization under weak supervision, and we can boost the generalization ability of our model by CDAM.**

***Index Terms*— Mitosis detection, box-supervised instance segmentation, domain adaptation, pesudo masks.**

Manuscript received 30 December 2021; revised 7 March 2022; accepted 29 March 2022. Date of publication 7 April 2022; date of current version 31 August 2022. This work was supported in part by NSFC under Grant 62103317; in part by the Shaanxi Natural Science Foundation under Grant 2022JM-335; in part by the Natural Science Foundation of Shaanxi Province under Grant 2021JQ-058; in part by the Science and Technology Program of Xi'an, China, under Grant 21RGZN0017; in part by the Beilin District Science and Technology Program under Grant GX2130; and in part by the Pazhou Laboratory, Guangzhou. *(Yonghui Li and Yao Xue contributed equally to this work.) (Corresponding author: Liangfu Li.)*

Yonghui Li and Yao Xue are with the School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: lyh1569843550@stu.xjtu.edu.cn; xueyao@xjtu.edu.cn).

Liangfu Li is with the School of Computer Science, Shaanxi Normal University, Xi'an 710119, China (e-mail: longford@xjtu.edu.cn).

Xingjun Zhang is with the Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: xjzhang@mail.xjtu.edu.cn).

Xueming Qian is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, and the SMILES Laboratory, School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/TMI.2022.3165518

## I. Introduction

MITOSIS detection is an important indicator of tumor cells in breast cancer diagnosis [1]. However, it is time-consuming and laborious for pathologists to manually complete the mitotic count. Therefore, it is necessary to develop automatic detection methods, which can not only save a lot of time, manpower and material resources, but also improve the reliability of pathological diagnosis [2].

Earlier mitosis detection approaches [3], [4] simulate the textural features and tissue morphology of mitotic cells to capture mitosis-specific characteristics for automated detection. However, due to the large intra-class variability of mitotic cells and the difficulty in distinguishing mitotic cells from normal cells, manual features are often poorly performing.

Recent CNN-based mitosis detection methods are mainly divided into three categories. (1) Mitosis detection by pixel classification [5], [6]. For each pixel of training images, it is labeled as mitotic when close to the centroid pixel of a mitotic cell, otherwise labeled as non-mitotic. Pixel classification is inherently a sliding-window-based method that produces a fixed-size patch for each pixel to be fed into the classification network, leading to high storage costs and inference time. (2)Mitosis detection by semantic segmentation [7], [8]. The semantic segmentation method directly predicts a segmentation map to determine the category of each pixel for the input image, avoiding duplicate computation. Pixel-level annotations are often required to train segmentation networks, so how to extend weak labels (i.e., only centroid coordinates of mitotic cells are provided) into precise semantic masks is particularly crucial. (3) Mitosis detection by object detection [9], [10]. Detection networks require bounding box annotations for training. While fortunately, it is relatively easy to generate ground truth boxes based on weak labels. This method has a significant improvement in detection speed.

The existing CNN-based mitosis detection methods still have some drawbacks. (1) It is difficult to give precise instance mask predictions for mitosis datasets with only weak labels. The instance mask predictions not only provide more information for disease diagnosis but also reduce the labeling work of pathologists significantly. To address this problem, Li *et al.* [7] designs a concentric loss function to train semantic segmentation networks with weak labels. But this method needs cumbersome filtering mechanisms and is hard to give precise pixel-wise mitosis localization. (2) Mitosis models

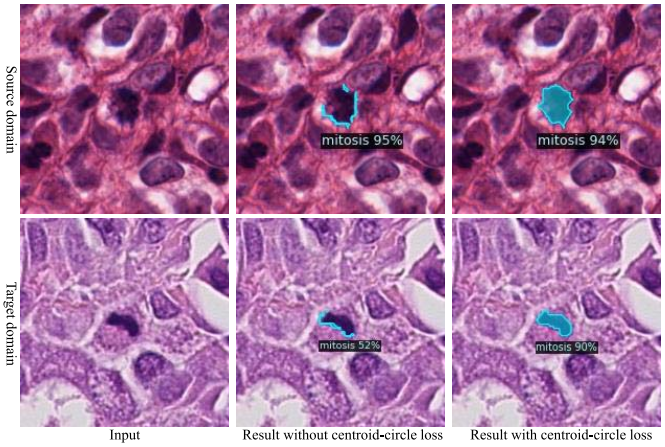



Fig. 1. The first row shows pathology images from $Lab_1$, while the second row corresponds to pathology images from $Lab_2$ and $Lab_3$. There is a significant domain gap between them. We also present some detection results of our model on these images: the second column shows the segmentation results without using centroid-circle loss in BIA head, and the third column shows the results with centroid-circle loss in BIA head.

learned from one domain may not be globally applicable. Due to the use of different microscope scanners and staining procedures, the histopathology images obtained from different institutions usually have high variability in appearance as shown in Fig.1. This hinders the generalization ability of the model, and the model trained in one pathology laboratory is difficult to use for the data of other laboratories. Various strategies have been proposed to address the domain migration issue, such as stain standardization [11], staining augmentation [6], and domain adversarial training for classification networks [12]. But these methods don't perform feature alignment at both the image and instance level, leading to much noise from background regions on the target domain.

We propose an approach to overcome the above challenges. In contrast to these methods, we argue that it makes more sense to consider mitosis detection as an application of instance segmentation combined with unsupervised domain adaptation. Therefore, we propose a Domain adaptive Box-supervised Instance segmentation Network (DBIN) for mitosis detection. DBIN can not only perform pixel-level mitosis prediction for weakly labeled mitosis datasets, but also adapt detectors trained on the source domain to an unlabeled target domain.

On the one hand, the major obstacle for the application of instance segmentation models is highly time-consuming pixel-level mask annotations. Several works [13]–[15] try to obtain instance masks through box annotations. PAD [13] and BBTP [14] are based on Faster-RCNN [16], and rely on region of interests(ROIs) on feature maps for further instance mask predictions. In addition, to obtain good performance, these methods need to first extract the object contours using traditional contour extraction algorithms (e.g., MCG [17] or GrabCut [18]), and then further refine the segmentation results by the iterative training process. These methods have complex network structures and require iterative training. In contrast, our approach is an anchor-free model, which not only eliminates the unfavorable deployment and time-consuming ROI operations but also does not require contour extraction algorithms that are difficult to compute in parallel on GPUs. In this paper, we create two novel branch networks for DBIN: a Box-supervised Instance-Aware (BIA) head for generating a single instance mask prediction for each mitotic cell; a Pseudo-Mask-supervised Semantic (PMS) head for enriching characteristic information extracted from feature maps. Specifically, instead of designing a tedious network structure for BIA head, we redesign three loss terms based on bounding box annotations according to some prior information of mitotic cells, including projection loss, consistency loss of adjacent pixel pairs, pixel-level centroid-circle loss. PMS head is used to assist in the learning of the instance segmentation task, which is supervised by pseudo ground truth masks generated by BIA head.

On the other hand, domain adaptation is necessary to be considered for mitosis detection. Since the model trained in one histopathological lab is hard to work well on unlabeled data from another lab, due to the significant discrepancy in data distribution. Various approaches [19]–[22] have been proposed by many researchers. To address the severe lack of target domain data and the problem of over-domain adaptation, Wang *et al.* [20] introduces a matching mechanism for source and target features, including adaptive components and feature regularization components. In the field of medical image processing, researches on domain adaptation are at the very beginning stage. Such works are mainly focused on classification tasks. Zhang *et al.* [19] proposes a DMAN network that uses adversarial learning to extract inter-domain consistency features. We argue that the domain adaptation problem for mitosis detection is a key limitation to the generalized application of models. One solution is to add constraints during the training stage so that the model ignores irrelevant changes in appearance. Facing the issue of domain gaps, we propose a Cross-Domain Adaptive Module (CDAM) for aligning feature distributions between domains. Specifically, we design two domain adaptive components in CDAM: a global adaptive component for image-level feature alignment, and a center-aware adaptive component that focuses on foreground region feature alignment. As shown in Fig.2, by gradually adding the global adaptive component and the center-aware adaptive component, we can enable our model response less to background regions (especially normal cells) in the target image.

The contributions of this paper can be summarized as:

(1) To train instance segmentation networks based on bounding box annotations, we propose a box-supervised instance-aware (BIA) head and give precise pixel-level mitosis predictions. Since instance masks can yield better localization than weak labels, mask predictions obtained from our model can be employed to boost the performance of other models without any additional annotations of ground truth masks.

(2) We create a pseudo-mask-supervised semantic (PMS) head whose purpose is to directly optimize the characteristics extracted from the feature space, ultimately facilitating the learning of detailed pixel-level information.

(3) To reduce domain gaps, we propose a cross-domain adaptive module (CDAM), including a global adaptive

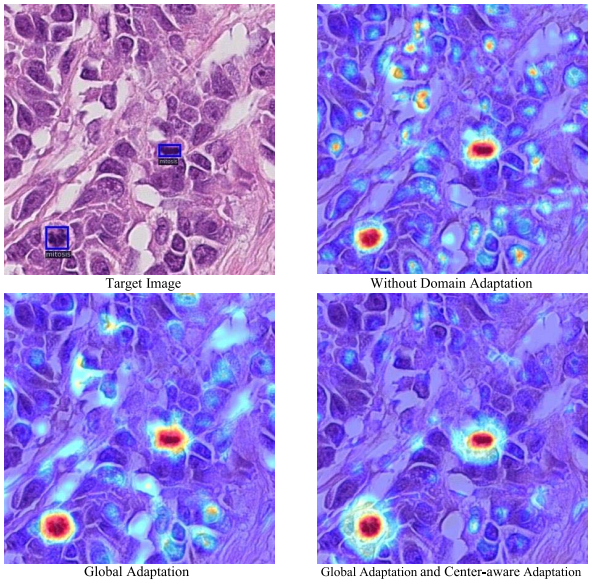



Fig. 2. Response maps on the target image. By gradually using the global adaptation and the center-aware adaptation in CDAM, we can reduce noise from background regions (especially normal cells).

component and a center-aware adaptive component, which align the feature distributions of both domains on the pixel level. In this way, we can learn a model adapted to unlabeled target domains, and we can apply our model in different laboratories and various pathological images.

(4) We design an end-to-end domain adaptive box-supervised instance segmentation network(DBIN) to localize mitosis in diverse histopathology images without cumbersome filtering mechanisms for model outputs. To our best knowledge, this is the first time that combining box-supervised instance segmentation with domain adaptation has been applied to mitosis detection. Also, We have achieved state-of-the-art results on four mainstream datasets with a fast detection speed.

## II. Related Work

### A. Mitosis Detection

Thanks to publicly available mitosis datasets [23]–[27], many methods have been proposed for automatic mitosis detection in histopathological slide images. In terms of extracting image features, we can classify them into two types, manual-based features, and CNN-based features. Earlier approaches [3], [4] used a variety of handcrafted features to describe the appearance of mitotic cells. The handcrafted features, containing shape, statistical and textural characteristics, are usually built on the expertise of pathologists. However, due to the diverse appearance of mitotic cells, it is difficult to characterize all mitotic cells very accurately and comprehensively with manually designed features.

Convolutional neural networks have revolutionized the field of computer vision [28]. Recently, many deep learning-based methods [19], [29]–[32] have been applied to medical image analysis tasks and have yielded results superior to traditional methods. Xue *et al.* [29] combines convolutional neural networks with compressed sensing in end-to-end training method for cell detection task. They convert a classification problem to a regression problem, the discretization of output space and inter-class imbalance are mitigated significantly.

CNN-based features allow automatic and more efficient learning of mitosis features. However, many methods fail to predict precise instance masks for weakly labeled mitosis datasets, such as TUPAC16 [26] dataset, while pixel-level mitosis predictions can provide more specific evidence for the medical diagnosis. Sohail *et al.* [33] uses MITOS12 dataset [25] with mask labels to train a label-refiner for weakly labelled mitosis figures, but this label-refiner may provide poor predictions for datasets from external pathology centers. Li *et al.* [7] proposes to extend single-pixel labels to region labels with concentric circles and designs a "concentric loss function" to train semantic networks. However, this method assumes that mitotic cells are round, resulting in the inability to give accurate pixel-wise localization for mitotic cells that are variable in shape. In contrast, our method makes no assumptions about the shape of mitotic cells, but instead utilizes a Box-supervised Instance-Aware (BIA) head with three box-supervised mask loss terms to guide training.

Furthermore, current methods ignore domain shift problems, resulting in poor performance when applied directly to different labs. Several strategies [6], [11], [12] have been proposed to address the domain migration issue for mitosis detection. Tellez *et al.* [6] utilizes staining augmentation techniques to automatically generate diverse training samples, which effectively improves the robustness of the model against different staining protocols. But this approach can hardly perfectly simulate the realistic staining variations and multi-scanner difference. Lafarge *et al.* [12] apply the DANN network [34] directly to mitosis detection and only align domain features at the image level, leading to poor feature alignment in mitotic regions. By contrast, we propose a Cross-Domain Adaptive Module (CDAM) that performs pixel-by-pixel alignment and spotlights high-confidence regions of mitosis.

### B. Box-Supervised Instance Segmentation

Mitosis datasets usually do not have mask-level annotations, so we desire to inherit advantages of instance segmentation methods through box-supervised training. We adopt CondInst [35] as the baseline network in this paper.

Recently, weakly supervised learning receives widespread attention in the field of deep learning [36]. Weakly supervised object detection with image-level annotations has been extensively investigated in the past few years [37], [38]. While weakly supervised instance segmentation with box-level annotations has not been much explored yet. Box-supervised instance segmentation can considerably reduce the labeling difficulty and predict pixel-level instance masks rather than labeled boxes. For example, SDI [39] need to extract object contours as original masks using traditional contour extraction algorithms MCG [17] or Grabcut [18], and then further iterate to refine instance masks. BBTP [14] views the box-supervised instance segmentation task as a multi-instance learning (MIL) problem and generates positive and negative bags based on

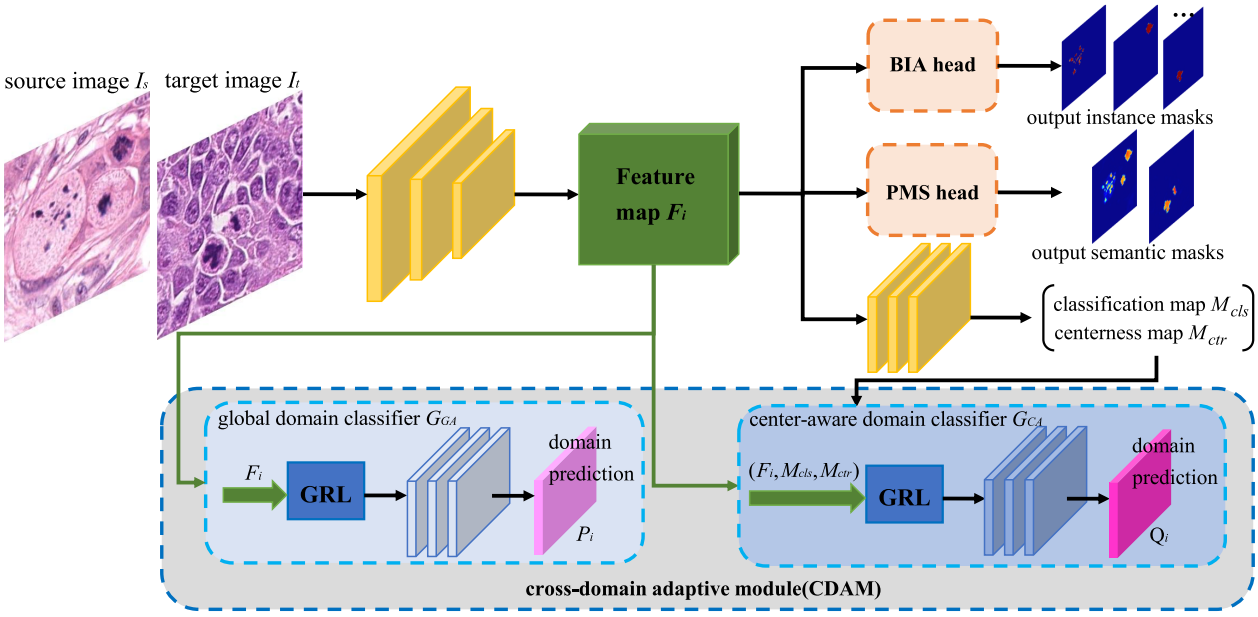



Fig. 3. System overview of the proposed Domain adaptive Box-supervised Instance segmentation Network (DBIN). DBIN contains three novel sub-networks: a Box-supervised Instance-Aware (BIA) head with the core idea of redesigning three box-supervised mask loss terms, which outputs instance masks for mitotic cells under weak supervision; a Pseudo-Mask-supervised Semantic (PMS) head that output semantic masks and enrich characteristic information extracted from underlying feature maps; and a Cross-Domain Adaptive Module (CDAM) that align feature distributions between domains. CDAM consists of two components: a global adaptive component for image-level feature alignment, and a center-aware adaptive component for mitosis-related feature alignment.

regional proposals. PAD [13] recursively estimates pseudo masks by an object detection branch and an instance segmentation branch. A top-down segmentation feedback is used to enhance its detection branch. In contrast, we propose a box-supervised instance-aware (BIA) head that contains three novel box-supervised mask loss terms instead of designing complex network structures, which not only facilitate training but also significantly improve the comprehensive performance, especially in terms of inference time.

### C. Domain Adaptation

Domain adaptation has been well explored in terms of image classification tasks [40], [41]. However, domain adaptation has been less studied in other computer vision tasks. To address domain adaptation problems in the field of object detection, several approaches [20]–[22], [42] have been proposed in recent years. For target domains with only image-level labels, Inoue *et al.* [42] uses a two-stage weakly supervised domain adaptation framework on the detector for fine-tuning. For target domains without annotations, Chen *et al.* [21] designs the image and instance level domain adaptation components based on Faster R-CNN for unsupervised adaptation. This method also enhances the consistency between two adaptation components through a consistency regularization component. In the community of medical image analysis, domain adaptation is now mostly used for classification tasks [12], [19], [43]. Zhang *et al.* [19] proposes a DMAN network to reduce domain differences through adversarial learning and entropy minimization. We observe that no domain adaptation model based on detection networks has been applied to the mitosis detection task. In this paper, we propose a cross-domain adaptive module (CDAM) that contains two domain adaptive components at the image and instance level, which are trained in an adversarial manner.

## III. Proposed Method

### A. Method Overview

Fig.3 shows an overview of our proposed framework. With three novel sub-networks, including a box-supervised instance-aware (BIA) head, a pseudo-mask-supervised semantic (PMS) head, and a cross-domain adaptive module(CDAM), we eventually form our domain adaptive box-supervised instance segmentation network (DBIN).

Our goal is to learn a box-supervised instance segmentation model adapted to the unlabeled target domain. We perform supervised training on source images $I_s$, but unsupervised training on target images $I_t$. The overall procedure is illustrated in Fig.3. Firstly, we extract the underlying feature map $F_i$ through feature extractors. Secondly, we use BIA head and PMS head to further extract instance characteristics from $F_i$ for pixel-level mitosis prediction. Lastly, we use CDAM to perform per-pixel domain predictions, which can learn domain-invariant feature representations for $F_i$.

### B. Box-Supervised Instance-Aware (BIA) Head

Here, we propose a high-performance box-supervised instance-aware (BIA) head that requires only bounding box annotations to implement instance segmentation. Instead of modifying the network structure for the original dynamic instance-aware mask head in the baseline network CondInst [35], our proposed BIA head simply redesigns three loss terms according to bounding box annotations.

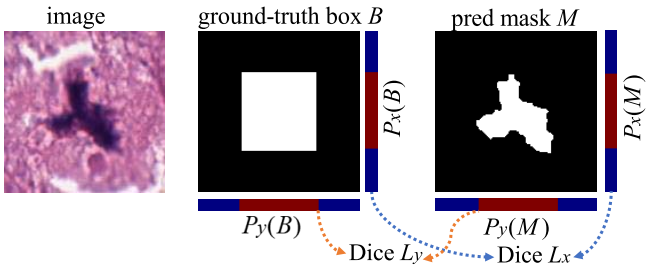



Fig. 4. Illustration of the projection loss in BIA head. We compute the projections of the ground-truth box and predection mask on the x- and y-axes, and use dice loss to overcome foreground sparsity.

*1) Projection Loss:* To enable the minimum bounding rectangle (MBR) of instance mask predictions match bounding box annotations, the projection loss is designed. As illustrated in Fig.4, for the horizontal and vertical projection vectors of the predicted mask, we evaluate the dice loss [44] separately.

Specifically, let $B \in \mathbb{R}^{H\times W}$ be supervised information generated on a whole training image by box annotations. Besides, the instance mask derived from the detector is represented as $M \in \mathbb{R}^{H\times W}$. We can calculate the horizontal projection vector $B_x \in \mathbb{R}^H$ and the vertical projection vector $B_y \in \mathbb{R}^W$ of $B$ via (1). Similarly, we can obtain $M_x$ and $M_y$ for $M$.

$$\begin{aligned} B_x(x) &= \max_y B(x,y) = \max\{B(x,y)|y = 1,2,\dots W\} \\ B_y(y) &= \max_x B(x,y) = \max\{B(x,y)|x = 1,2,\dots H\} \end{aligned} \quad (1)$$

To alleviate sample imbalances, we use the dice loss to shine a spotlight on foregrounds. The projection loss is defined as:

$$\begin{aligned} L_{proj} &= DL\left(\max_y B, \max_y M\right) + DL\left(\max_x B, \max_x M\right) \\ &= DL\left(B_x, M_x\right) + DL\left(B_y, M_y\right) \\ &= 2 - \frac{2\left|B_x \cap M_x\right|}{\left|B_x\right|^2 + \left|M_x\right|^2 + \varepsilon} - \frac{2\left|B_y \cap M_y\right|}{\left|B_y\right|^2 + \left|M_y\right|^2 + \varepsilon} \end{aligned} \quad (2)$$

where $DL$ function represents the dice loss [44], and $\varepsilon$ takes a default value of 1e-5 for the stability of denominator values.

*2) Consistency Loss of Adjacent Pixel Pairs:* For each location in the input image $I \in \mathbb{R}^{H\times W}$, there are eight adjacent pixels, namely the 8-neighbourhood. Under this definition, we can generate 8 neighbouring images $I_N^i$ $(i = 1,2,\dots 8)$ for image $I$, as shown in Fig.5. Observations reveal that the color inside mitotic cells is normally similar, but differs significantly from surrounding tissues. This prior knowledge makes color clustering feasible, and can be utilised to supervise training.

Firstly, we compute the color similarity map $S^i$ between the input image $I$ and its neighbouring image $I_N^i$ via (3),

$$S^i(x,y) = \exp(-\alpha\left|I(x,y) - I_N^i(x,y)\right|) \quad (3)$$

where $\alpha$ defaults to 2 in this paper. When the color similarity $S^i(x,y)$ is greater than the given threshold $\tau$ (defaulting to 0.3), we have reason to believe that the adjacent pixel pairs $(I(x,y),I_N^i(x,y))$ share an identical label; otherwise we do not make any assumptions. Based on this hypothesis, we produce the consistent supervision matrix $P^i$ for pixel pairs $(I,I_N^i)$, with $P^i(x,y)$ taking 1 only when $S^i(x,y)$ is greater than $\tau$, and 0 otherwise. As shown in Fig.5, white areas in $I_N^i$ are where $P^i$ takes 1.

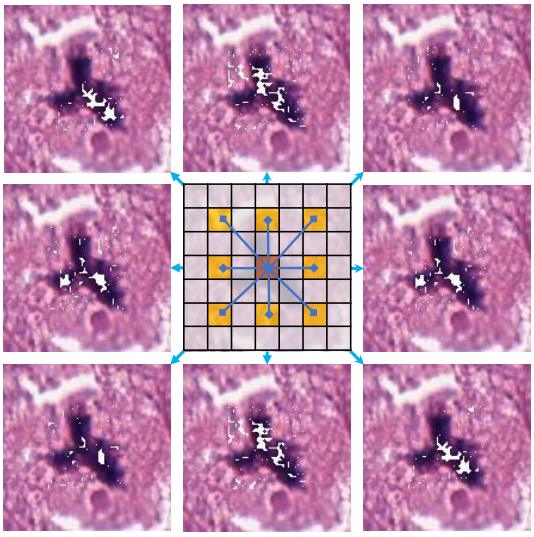


Fig. 5. Illustration of the consistency loss in BIA head. We calculate color similarities between every pixel and its 8-neighbourhood pixels. The white locations in 8 neighbouring images are regions where color similarities are over the threshold, and they are taken for model training.

Secondly, we generate 8 neighbouring instance masks $M_N^i$ $(i = 1,2,\dots 8)$ for the predicted instance mask $M$. Let the value of adjacent pixel pairs $(M(x,y),M_N^i(x,y))$ be $\hat{P}^i(x,y)$, whose value represents the prediction probability that adjacent pixel pairs have the same label and is calculated via (4).

$$\begin{aligned} \hat{P}^i(x,y) = M(x,y)\cdot M_N^i(x,y) \\ +(1 - M(x,y))\cdot\left(1 - M_N^i(x,y)\right) \end{aligned} \quad (4)$$

We only calculate the consistency loss for elements set to 1 in $P^i$, since pixel pairs at these positions are most likely to share the same label. Additionally, to avoid gradient updates being dominated by numerous pixel pairs outside ground-truth boxes, the consistency loss is calculated only for pixel pairs located inside boxes. Using the cross-entropy loss function to guide training, the consistency loss term is defined below:

$$L_{cons} = -\frac{1}{N}\sum_{i,x,y} B(x,y)P^i(x,y)\log(\hat{P}^i(x,y)) \quad (5)$$

where $N$ means the number of adjacent pixel pairs with consistent colours in the ground-truth box, and its value is expressed as $N = \sum_{i,x,y} B(x,y)P^i(x,y)$.

*3) Pixel-Level Centroid-Circle Loss:* By far, we have designed projection loss and consistency loss, which implement training box-supervised instance segmentation networks. However, because Eq.(5) computes losses only for colour-consistent pixel pairs inside bounding boxes, we observe that this may lead to several trivial solutions as shown in Fig.6(a). The first case is that all pixels of the predicted mask $M$ are 0, but this does not satisfy the projection loss term of Eq.(2). The second case is where $M$ degenerates to a bounding box, i.e., all pixels inside the box are 1, but the consistency loss tends to predict pixels

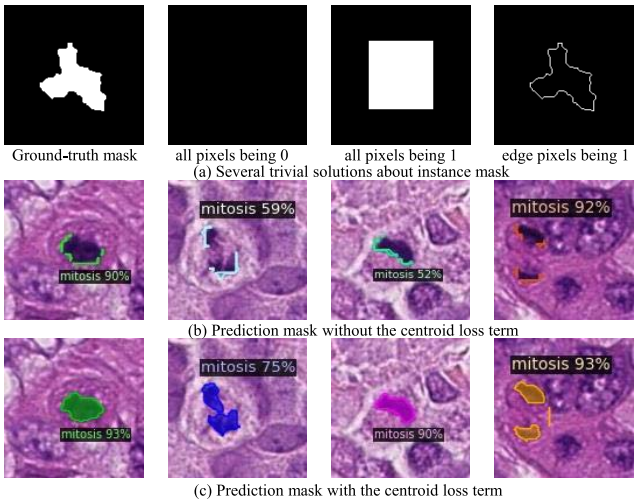



Fig. 6. Illustration of the centroid-circle loss in BIA head. (a) shows some trivial solutions about instance masks. (b) and (c) show the qualitative demonstration of the improvement by centroid-circle loss.

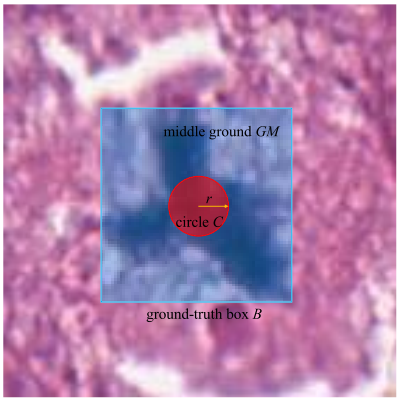



Fig. 7. Definition of the middle ground *GM*. *GM* is an area where mitotic and non-mitotic pixels coexist within the ground-truth box *B*.

on the edge of the box as negative samples. However, there is still a trivial solution that can't be avoided, where only pixels around cells' edges are 1. Not only does this satisfy the projection loss term, but the consistency loss term also falls into a local minimum solution. To prevent trapping into such a case shown in Fig.6(b), we propose a pixel-level centroid-circle loss, forcing center regions of mitosis to be positive.

As shown in Fig.7, We extend the centroid pixel label to a centroid-circle $C$ with an appropriate radius $r$ (defaulting to 5 pixels). $C$ is a high-confidence mitotic region, and the remaining area of the ground-truth box $B$ is defined as middle ground $GM$. Mathematically, $GM$ can be obtained via $GM = B - C$. $GM$ is an area where mitotic and non-mitotic pixels coexist within $B$, which is not involved in computing loss. We employ the dice loss function to supervise the predicted mask in the centroid-circle region as follows:

$$L_{cend} = \frac{2\sum \chi_{(x,y)\notin GM} M(x,y) \cdot B(x,y)}{\sum \chi_{(x,y)\notin GM} (M^2(x,y) + B^2(x,y))} \quad (6)$$

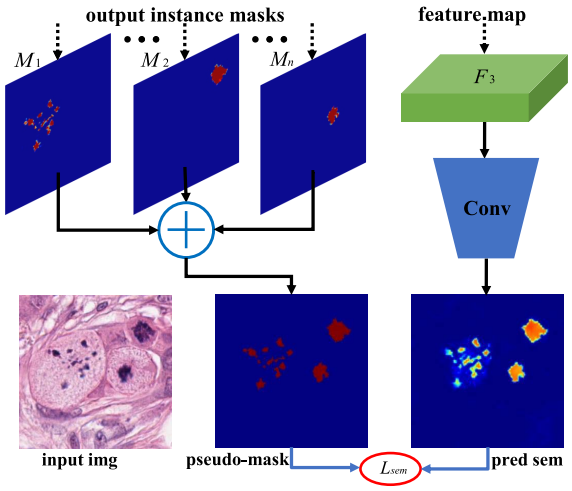



Fig. 8. The structure of pseudo-mask-supervised semantic (PMS) head. We generate a pseudo ground truth semantic mask using output instance masks from BIA head.

where $\chi$ denotes the indicative function. As shown in Fig.6(c), the above three loss terms can prevent falling into trivial solutions and yield well-formed instance masks.

### C. Pseudo-Mask-Supervised Semantic (PMS) Head

As indicated by YOLACT [45] and PAD [13], the object detection task and the instance segmentation task can potentially benefit from the joint semantic segmentation task. Therefore, we also designed a pseudo-mask-supervised semantic (PMS) head that shares the same feature maps with the BIA head. Instance-level detailed information obtained from PMS head can be back-propagated to enrich the detection characteristics.

As shown in Fig.8, we will fuse and binarize all instance mask predictions $M_i$ from BIA head to generate a pseudo ground truth semantic mask $M_s^{pse}$ via (7).

$$M_s^{pse}(x,y) = \text{sgn}(\max\{\sigma(M_i(x,y))|i=1,2,\ldots n\} - 0.5) \quad (7)$$

In this way, pseudo-masks and network parameters are alternately optimised for mutual gain. The focal loss ($FL$) is empolied to calculate the loss between $M_s$ and $M$ as below:

$$L_{sem} = FL(M_s, M_s^{pse}) \quad (8)$$

Additionally, to ensure training stability, we do not train PMS head in early epochs, as poor $M_s^{pse}$ leads to poor learning for segmentation, which in turn affects our framework adversely.

### D. Cross-Domain Adaptive Module (CDAM)

As illustrated in Fig.9, the cross-domain adaptive module (CDAM) is the last core component. CDAM is composed of a global adaptive component and a center-aware adaptive component. Notably, we align the pixel-wise feature distributions of two domains, which allows our model to take full account of every pixel and focus on the alignment of foreground pixels.

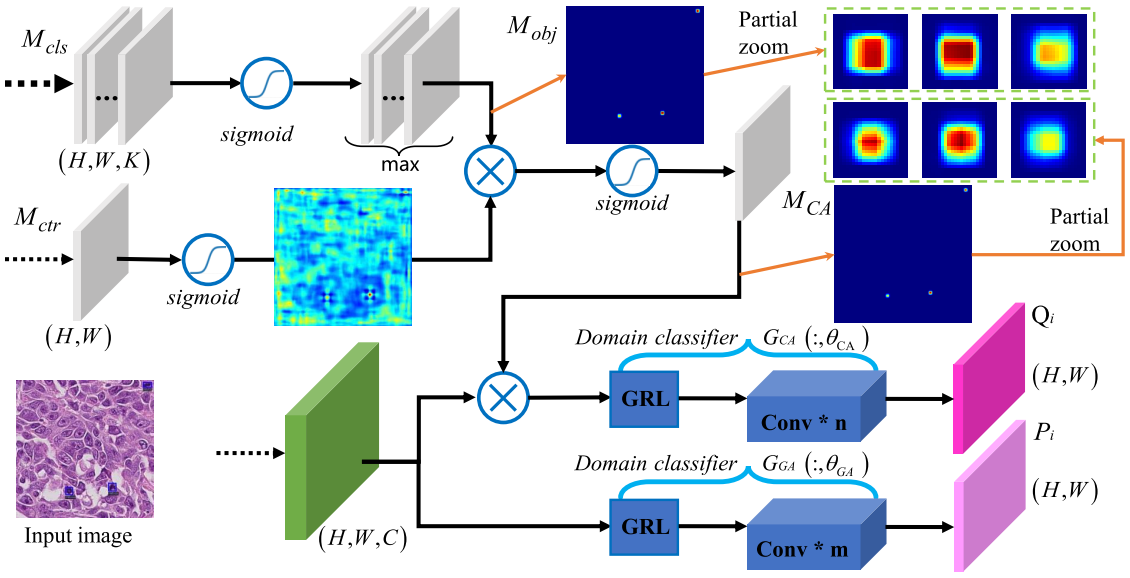



Fig. 9. The architecture of our cross-domain adaptive module (CDAM). It consists of two domain classifiers, including a global domain classifier $G_{GA}$ and a center-aware domain classifier $G_{CA}$. It is trained in an adversarial manner through a gradient inversion layer (GRL). The GRL ensures that the distribution of extracted features over the two domains becomes similar (as indistinguishable as possible for domain classifiers).

*1) Global Adaptive Component:* Image-level features are represented by feature maps extracted from the FPN at different levels. Specifically, we utilize five levels of feature maps, denoted as $\{F^3, F^4, F^5, F^6, F^7\}$. To align the domain distribution at the image level, we use a global domain classifier $G_{GA}$, as shown in the lower part of Fig.9.

We define $D_i$ to represent the domain label of the training image $I_i$, with $D_s = 1$ for the source image $I_s$ and $D_t = 0$ for the target image $I_t$. For the point $(u, v)$ of feature maps, we denote $P_i^{(u,v)}$ as the output of image $I_i$ obtained by $G_{GA}$. The cross-entropy loss function is used for $G_{GA}$, written as:

$$L_{GA} = -\sum_{u,v}\left[D_s \log P_s^{(u,v)} + (1-D_t)\log\left(1-P_t^{(u,v)}\right)\right] \quad (9)$$

In order to align the domain distribution, we use the gradient reversal layer (GRL) to achieve the ambition of inter-domain indistinguishability [34]. In this way, we can optimise the parameters of the domain classifier to minimise the domain classification loss via (9), together with optimising the parameters of our baseline network to maximise its loss. By adding the global adaptive component, we can reduce the performance degradation caused by pathological image differences.

*2) Center-Aware Adaptive Component:* Unlike approaches [20], [21], [42] based on region proposals to perform instance-level alignment, we adopt a center-aware adaptive component that allows us to focus on high-confidence regions of objects, which can reduce distractions from background regions.

In order to align the domain distribution at instance level, we train a domain classifier $G_{CA}$, as shown in the top half of Fig.9. Concretely, for the feature map $F_i$, we can obtain predictions of the classification map $M_{cls}$ and the centerness map $M_{ctr}$ from the detection head, and apply the sigmoid function to activate all elements. Then, we apply the max pooling operation to $M_{cls}$ to obtain the class-agnostic objectness map $M_{obj}$. $M_{obj}$ is then combined with the centerness map $M_{ctr}$ via (10) to obtain the center-aware map $M_{CA}$, giving particular interest to the central region of objects. To summarize, for a position $(u, v)$ at $M_{CA}$, its value is derived as follows:

$$\begin{aligned} M_{obj}^{(u,v)} &= \max_k\left(\sigma\left(M_{cls}^{(k,u,v)}\right)\right) \\ M_{CA}^{(u,v)} &= \sigma\left(M_{obj}^{(u,v)} \cdot \sigma\left(M_{ctr}^{(u,v)}\right)\right) \end{aligned} \quad (10)$$

We desire the center-aware domain classifier $G_{CA}$ to spotlight features in object regions from the feature map $F$. Therefore, we product the center-aware map $M_{CA}$ with $F$, which serves as the input of domain classifier $G_{CA}$.

Similar to the domain classifier $G_{GA}$, we insert the GRL layer into the front of domain classifier $G_{CA}$ so as to adopt the adversarial training technique. The output of image $I_i$ obtained by $G_{CA}$ at location $(u, v)$ is denoted as $Q_i^{(u,v)}$. Hence, the loss of our center-aware adaptive component can be written as:

$$L_{CA} = -\sum_{u,v}\left[D_s \log Q_s^{(u,v)} + (1-D_t)\log\left(1-Q_t^{(u,v)}\right)\right] \quad (11)$$

Aligning instance-level features can help reduce inter-domain distribution disparity among local instances. By adding a center-aware domain adaptive component, we provide the ability to focus on foreground regions, especially pixels with high objectness confidence and close to objects' centers.

## E. Overview of Loss

The total loss function of our DBIN is a weighted combination of components via (12), where $\lambda_i$ is the weight for balancing different loss terms. If not otherwise specified, we set $\lambda_1$ and $\lambda_3$ to 1 and $\lambda_2$ as 0.1 in this work.

$$\begin{aligned} L = Loss_{det} + \lambda_1(L_{proj} + L_{cons} + L_{cend}) \\ + \lambda_2 L_{sem} + \lambda_3(L_{GA} + L_{CA}) \end{aligned} \quad (12)$$

TABLE I
DETAILS OF TUPAC16 DATASET

| | $Lab_1$ | $Lab_2$ | $Lab_3$ |
|---|---|---|---|
| Case numbers | 1-23 | 24-48 | 49-73 |
| Origin | UMCUT | SPECA | SPECZ |
| Scanner | Aperio XT | Leica SCN400 | Leica SCN400 |
| Magnification | $40x$ | $40x$ | $40x$ |
| Spatial resolution | $0.25\mu m/pixel$ | $0.25\mu m/pixel$ | $0.25\mu m/pixel$ |
| Resolution | $2000 \times 2000$ | $5657 \times 5657$ | $5657 \times 5657$ |
| Area | $0.25mm^2$ | $2mm^2$ | $2mm^2$ |

TUPAC16 dataset consists of images from 73 breast cancer cases from three pathology centers. UMCUT denotes "University Medical Center in Utrecht", SPECA denotes "Symbiant Pathology Expert Center, Alkmaar", and SPECZ denotes " Symbiant Pathology Expert Center, Zaandam".

Notably, in the training phase, we feed the source domain data together with the target domain data into the network. During the testing phase, CDAM will be discarded and only cross-domain parameters will be used for mitosis detection.

Finally, we propose a solution for training a cross-domain instance segmentation model in a box-supervised way, which can significantly reduce the labeling work of pathologists.

## IV. EXPERIMENTS AND RESULTS

In this section, we evaluate the performance of our proposed domain adaptive box-supervised instance segmentation network (DBIN) on four mainstream datasets. We verify the effectiveness of our proposed BIA head, PMS head, and CDAM individually on TUPAC16 dataset [26].

We also perform extension analysis on various datasets. Firstly, we quantitatively evaluate the segmentation performance in MITOS12 dataset [25]. Besides, to further explore the Domain adaption capacity, we perform multi-dataset and multi-scanner cross-validation experiments. Moreover, we also compare the performance with more state-of-the-art methods to demonstrate the superiority of our method.

### A. Datasets

*1) TUPAC16 Dataset:* A full summary of TUPAC16 dataset [26] is provided in Table I. To facilitate training of the detection network, we take a few hours to manually expand weak labels into bounding box annotations that match the actual size and shape of mitotic cells. Notably, in a similar way to TUPAC16 dataset, we also extend the weak labels on MIDOG dataset [27] and MITOS14 dataset [24] into box annotations. The testing set contains images from 34 breast cancer cases. The mitosis annotations of this portion are not publicly available, and participants are evaluated by the contest organizers. However, this competition is currently closed, so we only validate on the training set.

We found that the images between three pathology laboratories have obvious differences in appearance and tissue texture. Especially these cases collected from $Lab_1$ truly differ from $Lab_2$ and $Lab_3$. Therefore, we used $Lab_1$ as the source domain dataset and $Lab_2$ and $Lab_3$ as the target domain dataset.

TABLE II
DETAILS OF MITOS12 DATASET

| Dataset | Scanner | Spatial resolution | Resolution | HPF numbers |
|---|---|---|---|---|
| Training | Aperio XT | $0.2456\mu m/pixel$ | $2084 \times 2084$ | 35 |
| | Hamamatsu | $0.2273\mu m/pixel$ | $2252 \times 2250$ | 35 |
| Testing | Aperio XT | $0.2456\mu m/pixel$ | $2084 \times 2084$ | 15 |
| | Hamamatsu | $0.2273\mu m/pixel$ | $2252 \times 2250$ | 15 |

TABLE III
DETAILS OF MITOS14 DATASET

| Dataset | Scanner | Spatial resolution | Resolution | HPF numbers |
|---|---|---|---|---|
| Training | Aperio XT | $0.2456\mu m/pixel$ | $1539 \times 1376$ | 1200 |
| | Hamamatsu | $0.2273\mu m/pixel$ | $1663 \times 1485$ | 1200 |
| Testing | Aperio XT | $0.2456\mu m/pixel$ | $1539 \times 1376$ | 496 |
| | Hamamatsu | $0.2273\mu m/pixel$ | $1663 \times 1485$ | 496 |

Following Li *et al.* [7] to perform dataset division, we take cases 30, 37, 44, 51, 58, 65, and 72 as the validation set, all of which are from the target domain, and the remaining as the training set. For the source domain, we always perform supervised training. While for the target domain, We will apply unsupervised training and test our model on the target domain.

*2) MITOS12 Dataset:* MITOS12 dataset [25] is extracted from 5 breast cancer biopsy slides. For each slide, an Aperio XT scanner and a Hamamatsu NanoZoomer scanner are used to produce one image respectively. A full summary of MITOS12 dataset is provided in Table II. There are 226 and 101 mitotic cells with mask annotations in the training and testing sets, respectively. Unless specifically stated, we use images produced by the Aperio XT scanner to evaluate our method.

*3) MITOS14 Dataset:* A full summary of MITOS14 dataset [24] is provided in Table III. There are 749 mitotic cells with only centroid annotations in the training sets. As mitotic annotations of the test set are not publicly available, we perform experiments by splitting the training set of MITOS14 dataset into training and validation data. Following the same splitting protocol in [9], [10], [46], we randomly divide the training set into training data and validation data by 4:1.

*4) MIDOG Dataset:* The MItosis DOmain Generalization (MIDOG) challenge [27] proposes a largest annotated multi-scanner and multi-center dataset on human breast cancer. The Training set consists of 200 images from four different scanners, with only 150 figure mitotic annotations being publicly available. A full summary of MIDOG dataset is provided in Table IV. Following [47], for each scanner, we divide it into training set and validation set in a ratio of 4:1.

*5) Performance Measurements:* According to the evaluation criteria, when the predicted position is smaller than a certain distance with the centroid of mitosis, this detection is correct. This threshold of distance is $5\mu m$ in MITOS12 dataset, but $7.5\mu m$ in the other datasets. We used $F_1$-score to evaluate the effectiveness of detection results.

TABLE IV
DETAILS OF MIDOG DATASET

| | $Scanner_1$ | $Scanner_2$ | $Scanner_3$ | $Scanner_4$ |
|---|---|---|---|---|
| Case numbers | $1-50$ | $51-100$ | $101-150$ | $151-200$ |
| Scanner | HAA XR | HAA S360 | Aperio | Leica |
| Spatial resolution | $0.23\mu m/px$ | $0.23\mu m/px$ | $0.25\mu m/px$ | $0.26\mu m/px$ |
| Resolution | $7215\times5412$ | $7094\times5370$ | $6475\times4840$ | $6413\times4655$ |
| Mitosis numbers | 451 | 582 | 688 | Unknown |

HAA denotes "Hamamatsu".

Besides, to evaluate the instance segmentation performance, we adopt the region-based measure $Pixel\ F_1$ as follows:

$$\begin{aligned} Pixel\ precision &= |PM \cap GM| \ / \ |PM| \\ Pixel\ recall &= |PM \cap GM| \ / \ |GM| \\ Pixel\ F_1 &= 2 \times \frac{(Pixel\ precision \times Pixel\ recall)}{(Pixel\ precision + Pixel\ recall)} \end{aligned} \tag{13}$$

where $PM$ is one instance segmentation mask prediction and $GM$ is the ground truth mask for a single mitotic cell. Thus, to calculate the $Pixel\ F_1$ score, we compute the averaged pixel precision for all multiple instance segmentation masks in an image. Additionally, we will also evaluate segmentation performance using the COCO-style mask AP, which is commonly used in the instance segmentation task.

### B. Hyper-Parameters

Our model with ResNet-101 as the backbone is initialized using weights pre-trained on ImageNet. For newly added layers, the parameters are initialized according to [35]. We firstly finetune the network with a learning rate of 1e-3 for 40k iterations and then continue for another 15k with a learning rate of 1e-4, and finally 15k with a learning rate of 1e-5. Each batch consists of 12 images, half from the source domain and the other half from the target domain. A weight decay of 1e-4 and a momentum of 0.9 are set up in our experiments.

It is worth noting that the consistency loss of adjacent pixel pairs in BIA head and semantic segmentation loss in PMS head are not involved in parameter updates at the beginning of training, and the loss weights are gradually increased from 0 starting at 10k and 30k iterations respectively, which can ensure the reliability of training and speed up the convergence.

### C. Data Augmentation of Training Data

We crop patches of $512 \times 512$ pixels with a step size of 128 pixels and re-scale them to $1024 \times 1024$ pixels. Furthermore, we use more data augmentation techniques to expand the training data (including random flip, elastic deformation, Gaussian blur, median blur, Gaussian noise, random lightning and contrast change, Random HSV, etc.). For source data, only patches containing mitotic cells will be used to train the detector. When it comes to target data, since it is trained in the unsupervised way, so the cropped patches are not filtered.

In the next few sections, we will gradually add our designed BIA head, PMS head and CDAM to the baseline and perform an experimental comparative analysis on TUPAC16 dataset.

TABLE V
PERFORMANCE OF DIFFERENT MASK LOSS TERMS IN BIA HEAD

| method | $L_{proj}$ | $L_{cons}$ | $L_{cend}$ | $F_1$-score |
|---|---|---|---|---|
| SegMitos [7] | | | | 0.710 |
| FCOS [48] | | | | 0.729 |
| Ours (BIA) | ✓ | | | 0.723 |
| | ✓ | ✓ | | 0.737 |
| | ✓ | ✓ | ✓ | **0.744** |

Results of BIA head with different mask loss terms on TUPAC16 dataset. "$L_{proj}$" means using projcetion loss, "$L_{cons}$" means using consistency loss, and "$L_{cend}$" means using centroid-circle loss.

### D. Influence of BIA Head

Firstly, we use a redesigned box-supervised instance-aware (BIA) head to replace original dynamical instance-aware mask heads in the baseline network. The training set (both source and target domains) generated in section IV-A.1 is fed into our designed box-supervised instance segmentation network, and the performance of our model is verified on the target domain.

We do not add PMS head and CDAM right now, and only analyze the impact of BIA head with three instance mask losses presented in section III-B with respect to our model performance. A comparison of performance between various models on TUPAC16 dataset is given in Table V.

Since our proposed instance segmentation system uses a similar detection module to FCOS [48], we focus on performance comparison with FCOS. To begin with, when using only the projection loss, the $F_1$-score of our model is 0.723, close to FCOS. Because the same projection vectors can correspond to numerous possible masks, the predicted mask is approximating to a box mask, leading to poor instance segmentation results. Secondly, when continuing to add the consistency loss of adjacent pixel pairs, a performance gain of $+1.4\%$ is achieved and the $F_1$-score reaches 0.737. By now, the edge and location features of objects can be nicely predicted by BIA head, as shown in Fig.6(b). However, it is observed that only pixels at the mitotic cell boundary are detected as positive samples, i.e., the predicted masks fall into a trival solution. Finally, we proceed to add the pixel-level centroid-circle loss term, forcing predictions around the object' centroid to 1, where a performance gain of $+2.1\%$ is implemented and the $F_1$-score reaches 0.744. With these three loss terms, we can train an instance segmentation network for mitotic cells based on box-level supervision only.

Table V primarily shows the advantages of our proposed model over other methods in terms of $F_1$-score. Nonetheless, our ambitions go beyond this, as we desire to make pixel-level predictions for mitotic cells. Fig.10 shows qualitative comparison results of different models in some sample regions of TUPAC16 dataset. Fig.10(a) indicates the bounding box annotations of all mitotic cells in the pathological images. Fig.10(b) shows the detection results of FCOS, and only box-level predictions are obtained. Fig.10(c) shows the instance segmentation results without using the centroid-circle loss term, and it can be seen that only pixels around cells' boundaries are predicted to be foreground regions. This indicates that relying

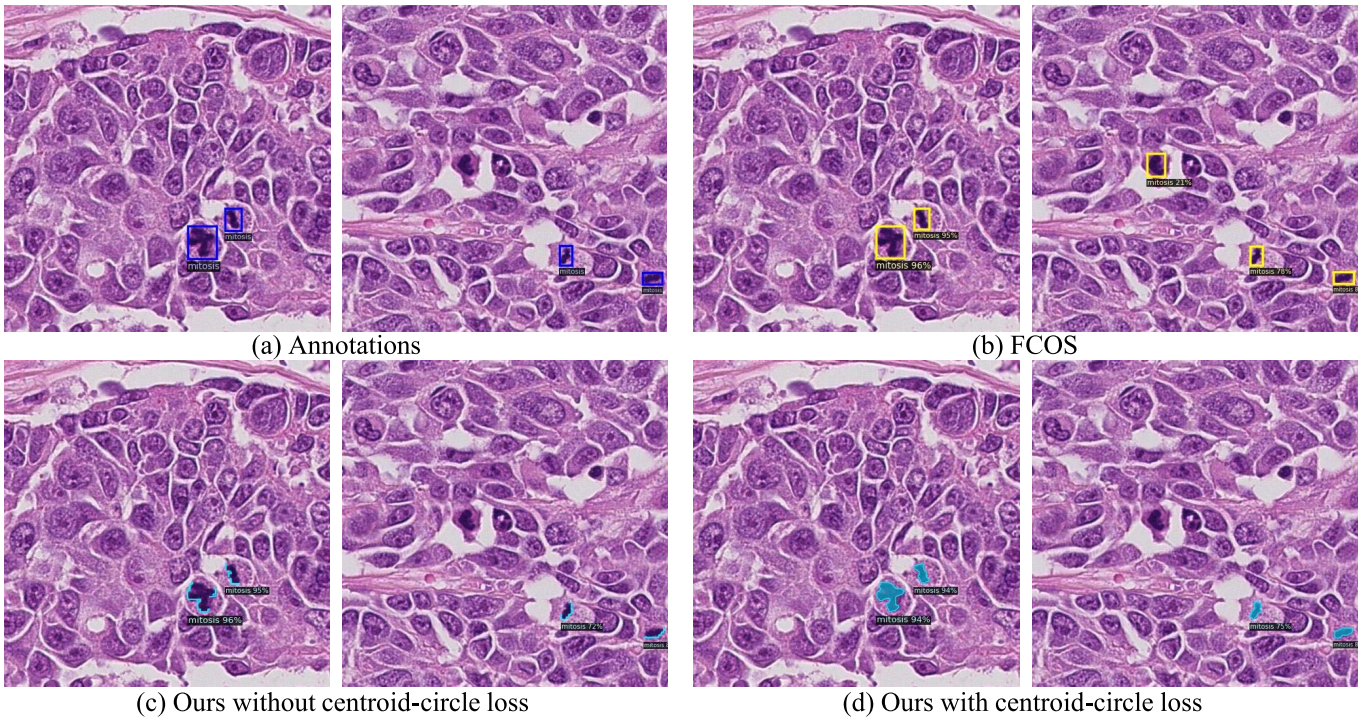



Fig. 10. Some detections comparisons between different methods. (a) shows bounding-box annotations of mitotic cells. (b) shows box-level detection results of FCOS [48]. (c) shows pixel-level instance predictions of our model without the centroid-circle loss term in BIA head, which fall into a trival solution. (d) shows precise mask predictions of our model with the centroid-circle loss term in BIA head.

on only two losses, projection loss and consistency loss, to supervise our network training is prone to yield solutions that do not match expectations. Finally, Fig.10(d) shows the segmentation images using three losses jointly involved in the training. In this case, the centroid region of the object tends to be 1, and it relies on the consistency loss for the foreground region growth, which ultimately yields accurate pixel-level predictions.

## E. Influence of PMS Head

Building on the model employed in the previous section, we go on to add a pseudo-mask-supervised semantic (PMS) head. BIA head can predict full-image instance masks, and we can stack these masks and binarize them to generate pseudo ground truth semantic masks. The flow of pseudo-masks' production and participation of training is shown in Fig.8.

The effects of PMS head on detection performance are given in Table VI. When using classification scores as NMS scores following FCOS [48], the $F_1$-score is 0.749, which is +0.5% higher than when there is no PMS head. Moreover, according to [49], the use of centerness and classification estimates are inconsistent during the training and inference phases, leading to certain performance degradation when directly multiplying the two as final classification scores during inference. An intuitive feeling is that the response of mitotic cells in the semantic mask should be large and bright, while the response of normal cells should be small and dark. Therefore, we consider using area-mean scores of a semantic mask as NMS score, which improves the $F_1$-score to 0.755. And, combining classification scores and area-mean scores as a ranking criterion, the performance continues to rise to 0.758. Finally, we performed test time augmentation (TTA) on the inference images, specifically, for each image, we rotate them with a 90-degree step and flip it. We calculate the average output of these variations, resulting in an $F_1$-score of 0.782.

TABLE VI
ABLATION EXPERIMENTS OF DIFFERENT TRICKS IN PMS HEAD

| method | cls | area | TTA | $F_1$-$score$ |
|---|---|---|---|---|
| Ours(BIA) | ✓ | | | 0.744 |
| Ours (BIA + PMS) | ✓ | | | 0.749 |
| | | ✓ | | 0.755 |
| | ✓ | ✓ | | 0.758 |
| | ✓ | ✓ | ✓ | **0.782** |

Ablation experiments of different tricks in PMS Head on TUPAC16 dataset. "cls" means using classification scores as NMS scores, "area" means using area-mean scores as NMS scores, and "TTA" means using test time augmentation.

## F. Influence of CDAM

Based on the previous two sections, we continue to add a cross-domain adaptive module (CDAM) for improving the cross-domain detection robustness of our model. In previous experiments, we use all annotations from the training set (both source and target domains). While, in this section, all experiments will apply unsupervised training to the target domain. We will also perform further analysis to show the impact of each domain adaptive component in CDAM.

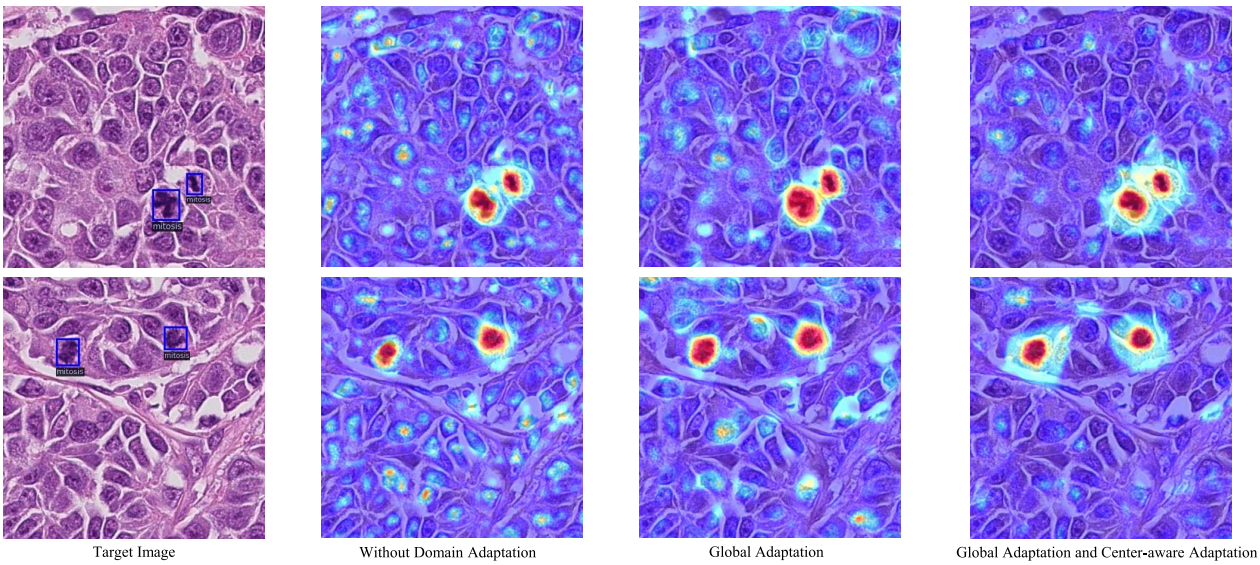



Fig. 11. Comparisons of response maps on target images. when the domain adaptation is disabled, our model will respond more to background regions(especially normal cells) in the target image. With the gradual addition of the global adaptive component and the center-aware adaptive component, our model can focus more on mitosis and reduce background noise.

TABLE VII
PERFORMANCE OF DIFFERENT COMPONENTS IN CDAM

| method | precision | Recall | $F_1$-$score$ |
|---|---|---|---|
| Ours w/o CDAM | 0.673 | 0.733 | 0.702 |
| Ours w/ CDAM (GA) | 0.753 | 0.677 | 0.713 |
| Ours w/ CDAM (CA) | 0.808 | 0.655 | 0.724 |
| Ours w/ CDAM (GA+CA) | 0.813 | 0.678 | **0.739** |

Performance Results of CDAM with different domain components on target domain (validation data of TUPAC16 dataset). Global adaptive and center-aware adaptive components are denoted as "GA" and "CA", respectively. We apply unsupervised training to the target domain ($Lab_2$ and $Lab_3$), and use only the supervised information from the source domain ($Lab_1$).

The effects of different domain adaptive components in CDAM for the detection performance on the target domain are provided in Table VII. Results show that each component in CDAM provides a performance gain to the detector compared to the $F_1$-score of 0.702 for the network "Ours w/o CDAM". In detail, we achieved a performance improvement of $+1.1\%$ using only image-level global feature alignment and $+2.2\%$ by adding only center-aware instance-level feature alignment. Also, applying both components at the image and instance level simultaneously yields a $3.7\%$ improvement.

Although the global and center-aware domain classifiers in CDAM focus on different tasks, they both guide the model to learn domain-invariant feature maps from different perspectives. For the global domain classifier, it focuses equally on each activation element in feature maps and determines whether it is from the source or target domain, which helps to initially reduce the image style gap. However, its training process may be dominated by numerous background pixels, resulting in poor feature alignment in the target region. To address this issue, we propose a center-aware domain classifier that achieves alignment of instance features by weighting the feature map to focus more on pixels that may be foreground.

In addition, we also perform a qualitative analysis. We present the response maps of several methods on target images used to localize mitotic cells. In Fig.11, when both domain adaptive components are disabled, the network "Ours w/o CDAM" will respond to normal cells in the target image. when the global adaptive component is enabled, the response map responds weaker to background regions, which can slightly reduce domain gaps due to image style differences. Further, when both our global and center-aware adaptive components are enabled, it can be observed that the response to the background is almost non-existent, which further reduces the inter-domain distribution differences between local instances. Therefore, adding domain adaptive components allows our model to focus more on objects and reduce the response to background noise.

### G. Comparison With Other Methods on TUPAC16

*1) Performance Comparison:* Finally, we achieved an $F_1$-score of 0.782 on TUPAC16 dataset, which outperformes the detection performance of all teams. And, we do not use any other dataset for training. As shown in Table VIII, the detection performance of various methods is provided. Specifically, when we use only source data for supervised training and verify the performance on target images, our method "Ours w/ CDAM" still outperforms other methods with an $F_1$-score of 0.739. This indicates that our domain adaptive box-supervised instance segmentation network (DBIN) can be directly applied to unlabeled pathology images from different laboratories while maintaining a good detection performance.

*2) Time Analysis:* The detection speed is a critical factor for clinical application. For the $2048 \times 2048$ pixels image in TUPAC 2016 dataset, our model's detection time is 1.29 seconds, and the GPU we used in all experiments

TABLE VIII
PERFORMANCE COMPARISON ON TUPAC16 DATASET

| Team | precision | Recall | $F_1$-$score$ |
|---|---|---|---|
| HUST, China | 0.640 | 0.700 | 0.669 |
| Lunit Inc, Korea | – | – | 0.652 |
| IBM Research Zurich and Brazil | – | – | 0.648 |
| Contextvision, Sweden | – | – | 0.616 |
| The Chinese University of Hong Kong | – | – | 0.601 |
| Microsoft Research Asia, China | – | – | 0.596 |
| Radboud UMC, The Netherlands | – | – | 0.541 |
| University of Heidelberg, Germany | – | – | 0.481 |
| University of South Florida, USA | – | – | 0.440 |
| Pakistan Institute of Engineering | – | – | 0.424 |
| University of Warwick, UK | – | – | 0.396 |
| Shiraz University of Technology, Iran | – | – | 0.330 |
| **SegMitos* [7]** | – | – | 0.710 |
| **FCOS* [48]** | 0.775 | 0.689 | 0.729 |
| **Ours*** | 0.787 | 0.778 | **0.782** |
| **Ours w/ CDAM* (only using source labels)** | 0.813 | 0.678 | **0.739** |

Performance of different approaches on TUPAC16 Dataset, whereas (*) indicates the same splitting protocol with our model. "HUST" denotes "Huazhong University of Science and Technology". "Ours w/ CDAM" denotes our model with a cross-domain adaptive module (CDAM), which uses only the source domain data $Lab_1$, i.e., the first 23 cases of the TUPAC16 dataset, for supervised training.

TABLE IX
SEGMENTATION PERFORMANCE COMPARISON ON MITOS12 DATASET

| Method | Pixel Precision | Pixel Recall | Pixel $F_1$ | Mask AP |
|---|---|---|---|---|
| Condinst [35] | 0.747 | 0.844 | **0.793** | **55.804** |
| Boxinst [15] | 0.706 | 0.808 | 0.754 | 51.334 |
| Ours (BIA+PMS) | 0.761 | 0.777 | **0.769** | **52.168** |

for calculating speed is one NVIDIA GeForce GTX 2080Ti. The SegMitos method [7], which currently performs best in mitosis detection, takes 3 seconds to detect one image. Compared to this, our method provides faster detection and better performance.

### H. Extensions: Evaluation of Segmentation Performance

To evaluate the segmentation performance, we conduct a comparison analysis on MITOS12 dataset that carries mask annotations.

*1) Comparison With SOTA Instance Segmentation Methonds:* Table IX presents the segmentation performance comparison of our method with SOTA box-supervised instance segmentation method BoxInst [15] and instance segmentation method CondInst [35]. "Ours (BIA + PMS)" outperforms the current SOTA box-supervised method BoxInst, but CondInst achieves the best performance because it uses original mask annotations for training (as pixel-level masks contain more precise localization than box-level annotations). However, our model shows better cross-domain robustness than CondInst and BoxInst in that it restrains the performance degradation caused by domain shifts.

*2) Domain Adaption Capacity Comparison With SOTA Instance Segmentation Methonds:* The domain adaptation results of different models are reported in Table X. For each method, the first row provides the performance of the source domain (i.e., validation images created by scanner Aperio XT), while the second row presents the performance of the target domain (i.e., validation images created by scanner Hamamatsu). CondInst yields the best segmentation performance, with a $Pixel\ F_1$ score of 0.793 and mask AP of 55.8% in the source domain. But there is a strong performance degradation on the target domain, resulting in a $Pixel\ F_1$ score of 0.701 and mask AP of 45.670% for CondInst. While the inclusion of CDAM allows our model "Ours w/ CDAM" to obtain the best detection and segmentation performance on the target domain, with an of 0.816, $Pixel\ F_1$ score of 0.737, and mask AP of 48.500%. It indicates that by adding domain adaptive components, we can generate a robust detector for the unlabeled target domain.

TABLE X
DOMAIN ADAPTION CAPACITY COMPARISON ON MITOS12 DATASET

| Method | Validation domain | $F_1$-$score$ | Pixel $F_1$ | Mask AP |
|---|---|---|---|---|
| Condinst [35] | Aperio XT (source) | **0.832** | **0.793** | **55.804** |
| | Hamamatsu (target) | 0.786 | 0.701 | 45.670 |
| BoxInst [15] | Aperio XT (source) | 0.812 | 0.754 | 51.334 |
| | Hamamatsu (target) | 0.759 | 0.677 | 39.978 |
| Ours w/o CDAM | Aperio XT (source) | 0.825 | 0.769 | 52.168 |
| | Hamamatsu (target) | 0.775 | 0.694 | 41.700 |
| Ours w/ CDAM | Aperio XT (source) | 0.824 | 0.785 | 52.610 |
| | Hamamatsu (target) | **0.816** | **0.737** | **48.500** |

TABLE XI
RESULTS ON THE TEST SET OF MITOS12 DATASET

| Method | Precision | Recall | $F_1$-$score$ |
|---|---|---|---|
| IPAL [25] | 0.698 | 0.74 | 0.718 |
| HC+CNN [50] | 0.84 | 0.65 | 0.735 |
| IDSIA [5] | 0.88 | 0.70 | 0.782 |
| CasNN [8] | 0.804 | 0.772 | 0.788 |
| DeepMitosis [9] | 0.846 | 0.762 | 0.802 |
| RRF [51] | 0.835 | 0.811 | 0.823 |
| SegMitos [7] | 0.854 | 0.812 | **0.832** |
| Condinst [35] | 0.832 | 0.832 | **0.832** |
| Boxinst [15] | 0.833 | 0.792 | 0.812 |
| Ours (BIA + PMS) | 0.810 | 0.842 | **0.825** |

*3) Comparison With SOTA Mitosis Detection Methods on MITOS12 Dataset:* Finally, Table XI presents the results of our method and current mitosis detection methods. Using only box-level labels, our method achieves a top-2 performance, and the $F_1$-score reaches 0.825.

### I. Extensions: Explore the Domain Adaption Capacity

*1) Multi-Dataset Cross-Validation:* To further prove that our proposed method can significantly alleviate the domain migration problem due to multiple centers and multiple scanners. We further explore the generalization performance of our model among TUPAC16 dataset, MITOS12 dataset, and MITOS14 dataset. The multi-dataset cross-validation results

TABLE XII
MULTI-DATASET CROSS-VALIDATION RESULTS AMONG TUPAC16, MITOS12 AND MITOS14 DATASETS

| Source data | Target data | Validation data | $F_1$ w/o CDAM | $F_1$ w/ CDAM |
|---|---|---|---|---|
| TUPAC16 $Lab_1$ | TUPAC16 $Lab_{2,3}$ | TUPAC16 $Lab_{2,3}$ | 0.702 | **0.739** |
| | | M12 Hamamatsu | 0.633 | **0.651** |
| | | M14 Hamamatsu | 0.543 | **0.592** |
| M12 Aperio | M12 Hamamatsu | M12 Hamamatsu | 0.775 | **0.816** |
| | | M14 Hamamatsu | 0.408 | **0.449** |
| | | TUPAC16 $Lab_{2,3}$ | 0.492 | **0.542** |
| M14 Aperio | M14 Hamamatsu | M14 Hamamatsu | 0.620 | **0.654** |
| | | TUPAC16 $Lab_{2,3}$ | **0.567** | 0.563 |
| | | M12 Hamamatsu | 0.678 | **0.695** |

"M12", "M14" denote the MITOS12 dataset and MITOS14 dataset, respectively. "Aperio" and "Hamamatsu" are two different scanners.

TABLE XIII
MULTI-SCANNER CROSS-VALIDATION RESULTS ON MODOG DATASET

| Source data | Target data | Validation data | $F_1$ w/o CDAM | $F_1$ w/ CDAM |
|---|---|---|---|---|
| $Scanner_1$ | $Scanner_2$ | $Scanner_1$ | **0.753** | 0.747 |
| | | $Scanner_2$ | 0.643 | **0.699** |
| | | $Scanner_3$ | 0.720 | **0.751** |
| $Scanner_2$ | $Scanner_3$ | $Scanner_2$ | **0.778** | 0.773 |
| | | $Scanner_3$ | 0.750 | **0.827** |
| | | $Scanner_1$ | 0.669 | **0.709** |
| $Scanner_3$ | $Scanner_1$ | $Scanner_3$ | 0.829 | **0.852** |
| | | $Scanner_1$ | 0.691 | **0.718** |
| | | $Scanner_2$ | 0.576 | **0.656** |

are reported in Table XII. The first column shows source domain data used to train our model, and the second column presents target domain data used to apply unsupervised training. For each dataset, we provide the performance of the internal target domain in the first row, and the second row and third row present the detection performance of external target domains.

Table XII compares the results with and without CDAM among various datasets. For the model trained on TUPAC16 dataset, while testing the performance on the internal target domain TUPAC16 $Lab_{2,3}$, we achieve an $F_1$-score of 0.739 with CDAM, resulting in a performance improvement of $+3.7\%$ compared with 0.702. As expected, the $F_1$-scores on external target domains MITOS12 Hamamatsu and MITOS14 Hamamatsu with CDAM are improved to 0.651 and 0.592, from 0.633 and 0.543, respectively, which are due to CDAM's strong domain adaption capacity. With the adoption of CDAM, for these three datasets, the validation results on the internal target domain and the other two external target domains are almost all improved. Thus, CDAM can bring better generalization performance for both internal and external target domains.

*2) Multi-Scanner Cross-Validation:* There are certain differences in manual mitosis diagnostic criteria among various datasets, so we also perform a comparison analysis on the multi-scanner MIDOG dataset. Multi-scanner cross-validation results on MIDOG dataset are reported in Table XIII. For each model, we provide the performance of the source domain, internal target domain, and external target domain in each row sequentially. Specifically, for the model using $Scanner_1$ as source data, the $F_1$-score of 0.747 on the validation data $Scanner_1$ (from source domain) with CDAM remains basically unchanged compared with 0.753, which is because the source domain and validation domain have no discrepancy of data distributions. While testing the performance on $Scanner_2$ (internal target domain), we achieve an $F_1$-score of 0.699 with CDAM, resulting in a performance improvement of $+5.6\%$ compared with 0.643. Besides, as expected, our model with CDAM also yields an improvement of $+3.1\%$ on $Scanner_3$ (external target domain), and the $F_1$-score is improved to 0.751 from 0.720. Similarly, we also carry out two other sets of experiments. With the inclusion of CDAM, for these three scanners, the validation performances on both the internal and external target domains are significantly improved.

TABLE XIV
K-FOLD CROSS VALIDATION PERFORMANCE ON TUPAC16 DATASET

| Fold | Precision | Recall | $F_1$-*score* |
|---|---|---|---|
| $K_1$ | 0.772 | 0.888 | 0.826 |
| $K_2$ | 0.821 | 0.674 | 0.740 |
| $K_3$ | 0.788 | 0.839 | 0.812 |
| $K_4$ | 0.711 | 0.818 | 0.761 |
| $K_5$ | 0.735 | 0.781 | 0.758 |
| $K_6$ | 0.569 | 0.707 | 0.630 |
| $K_7$ | 0.787 | 0.778 | 0.782 |

TABLE XV
METHODS PERFORMANCE ON MITOS14 DATASET

| Method | Precision | Recall | $F_1$-*score* |
|---|---|---|---|
| DeepMitosis* [9] | – | – | 0.572 |
| Cai et al.* [46] | 0.53 | 0.66 | 0.585 |
| SegMitos [7] | 0.495 | 0.785 | 0.607 |
| Mahmood et al.* [10] | 0.848 | 0.583 | **0.691** |
| Ours (BIA+PMS)* | 0.581 | 0.691 | **0.631** |

* Indicates the same splitting protocol with our model.

## J. Extensions: K-Fold Cross Validation on TUPAC16

To demonstrate the reliability and repeatability of experimental results when using other cases as validation. We divide cases 24-73 into training and validation set with a 6:1 ratio, and perform 7 splits totally, namely $K_i (i = 1, 2, \ldots 7)$. For each split $K_i$, we select cases in $[23+i, 30+i, 37+i, 44+i, 51+i, 58+i, 65+i]$ as the validation data. Specifically, the split $K_7$ is our old splitting protocol. Table XIV presents detection performances of "ours (BIA+PMS)" in various splits. Through k-fold cross validation, we obtain a mean $F_1$-score of 0.758 with a standard deviation of 0.059. It indicates that our model is robust and reliable.

## K. Extensions: Comparison Analysis With More Methods

*1) Performance Comparison on MITOS14 Dataset:* We compare the performance with DeepMitosis [9] and

TABLE XVI
PERFORMANCE COMPARISON ON MIDOG DATASET

| Method | Precision | Recall | $F_1$-*score* |
|---|---|---|---|
| Reference Algorithm [47] | 0.694 | 0.819 | 0.751 |
| Ours (BIA+PMS) | **0.747** | **0.854** | **0.797** |

TABLE XVII
PERFORMANCE COMPARISON WITH MP-MITDET [33]

| Method | Precision | Recall | $F_1$-*score* |
|---|---|---|---|
| Wahab et al. [52] | 0.77 | 0.66 | 0.713 |
| MP-MitDet [33] | 0.734 | 0.768 | 0.750 |
| Ours (BIA+PMS) | 0.735 | 0.815 | **0.773** |

Mahmood *et al.* [10] on MITOS14 dataset. We use the same splitting protocol mentioned in [9], [10], [46] for obtaining a fair comparison. Results of our proposed method and state-of-the-art methods are provided in Table XV. Our method performs better than other models except for Mahmood *et al.* [10], achieving a top-2 performance. However, Mahmood *et al.* [10] proposes a multi-stage mitosis detection method "FRCNN + PP + SF", making the computational cost extremely expensive. By contrast, our method not only has a faster detection speed with fewer computational resources, but also boosts the domain generalization ability of mitosis detection.

*2) Performance Comparison on MIDOG Dataset:* Table XVI shows the performance of various models on the validation data of MIDOG dataset. The $F_1$-score of our model is 0.797, which achieves a +4.6% performance improvement over the reference algorithm [47] of the MIDOG Challenge.

*3) Performance Comparison With MP-MitDet [33]:* To get comparable results with MP-MitDet [33], we report the performance of $F_1$-score on challenging TUPAC16 dataset on Table XVII. We use the same cross-validation scheme as mentioned in [33], [52], where patient samples are kept independent during training, validation and testing to simulate real-world situations. Our model performs best in discrimination of mitoses with an $F_1$-score of 0.773 on the test set.

*4) Computational Complexity:* Apart from time analysis with state-of-the-art mitosis detection models, we also measure the computational complexity of our model and other state-of-the-art box-supervised instance segmentation models. We use three common metrics: (1) number of model parameters; (2) FLOPS (floating-point operations per second); (3) the average inference time for one input image ($1024 \times 1024$ pixels). Table XVIII summarizes the comparison results with the three state-of-the-art methods: including PAD [13], BBTP [14], and BoxInst [15]. PAD comprises of an object detection branch and an instance segmentation branch, with a total of 134M parameters and 304.1G FLOPS, and such structure also slows down its speed of inference by 490ms per image. BoxInst and our model DBIN are both built on FCOS [48], making them have the similar number of parameters (30.26M vs 32.52M) and speed (82ms vs 80ms). However, our method has several additional parallelized auxiliary branches

TABLE XVIII
COMPARISON ON COMPUTATIONAL COMPLEXITY WITH BOX-SUPERVISED INSTANCE SEGMENTATION METHODS

| Methods | Parameters | FLOPS | inference time |
|---|---|---|---|
| PAD [13] | 61.6M | 304.1G | 490 ms |
| BBTP [14] | 65.3M | 332.6G | 215 ms |
| BoxInst [15] | **30.3**M | **137.6**G | **80** ms |
| Ours | **32.5**M | **152.0**G | **82** ms |

TABLE XIX
PERFORMANCE COMPARISON WITH SOTA OBJECT DETECTION METHODS ON TUPAC16 DATASET

| Method | Precision | Recall | $F_1$-*score* |
|---|---|---|---|
| ATSS [53] | 0.683 | 0.767 | 0.723 |
| GFL [49] | 0.734 | 0.757 | 0.745 |
| FCOS [48] | 0.775 | 0.689 | 0.729 |
| Ours(BIA+PMS) | 0.787 | 0.778 | **0.782** |

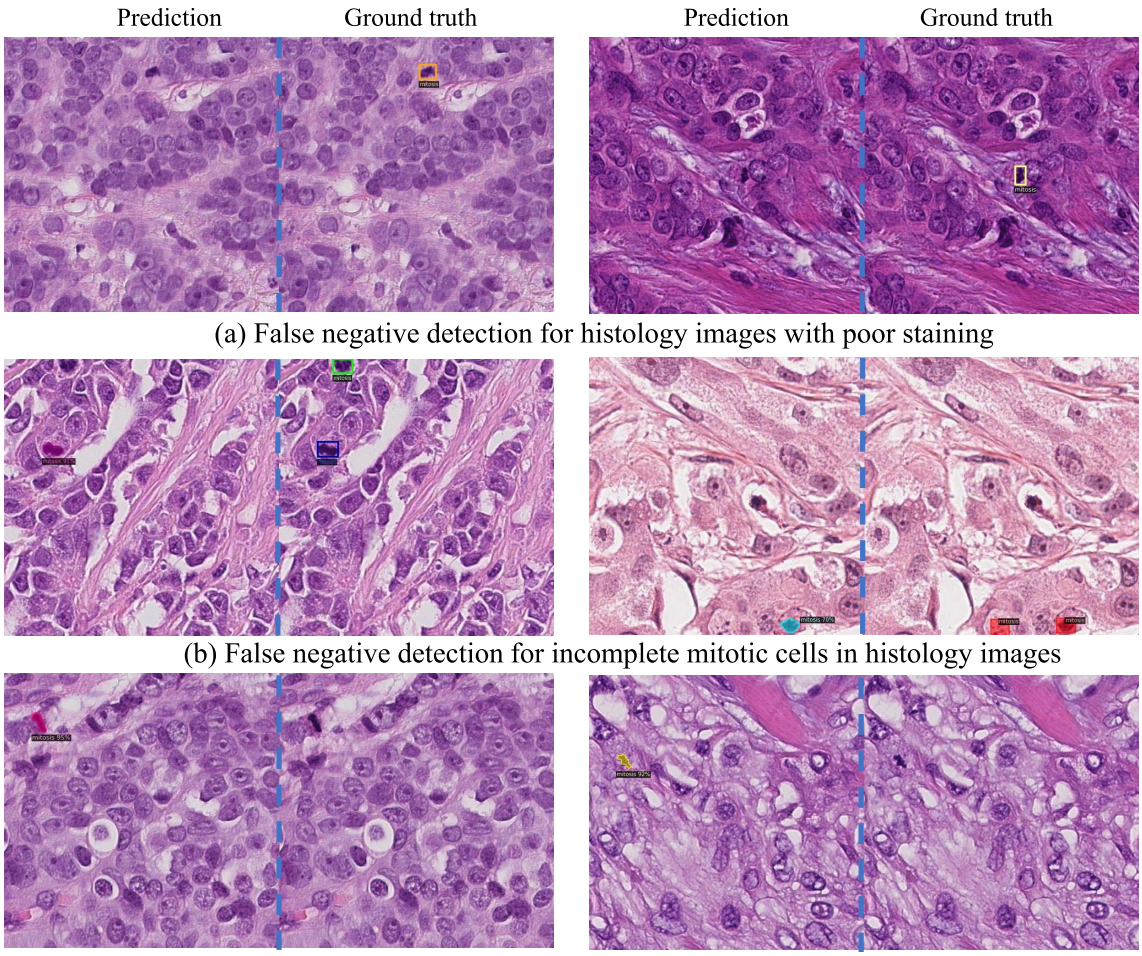
(e.g., the PMS head and BIA head), making our method a little more complex. Overall, compared with current box-supervised instance segmentation methods, our model has a relatively low computational complexity.

*5) Performance Comparison With SOTA Object Detection Methods:* In our experiments, FCOS [48] achvies an $F_1$-score of 0.729 on TUPAC16 dataset, better than other methods (except our proposed method). Mitotic detection is commonly regarded as an object detection task due to the powerful feature extraction capability of current deep CNN detection models. Besides, the following tricks in our experiments can bring additional benefits for the excellent performance of FCOS for mitosis detection: (1) More data augmentation techniques; (2) Multi-scale training; (3) We manually extend the weak labels into box annotations. Moreover, we conduct a comparative analysis with three state-of-the-art methods ATSS [53], GFL [49], and FCOS [48]. Table XIX summarizes the comparison results on TUPAC16 dataset. Thus, I think using object detection methods with our proposed dedicated method improvements based on the characteristics of pathology images is key for mitosis cell location task.

## L. Extensions: Ablation Study

We carry out ablation experiments to better understand some hyper-parameters in our model.

*1) Hyper-Parameters of the Consistency Loss Term:* To investigate that how the color similarity threshold $\tau$ in the consistency loss term affects the segmentation performance, we conduct experiments on MITOS12 dataset. We also compare with the SOTA instance segmentation method CondInst [35]. As shown in Table XX, when we set $\tau$ to 0, all adjacent pixel pairs will be involved in training. In this case, it is equivalent to assuming that all neighboring pairs have the same label, which leads to a large number of elements in the consistent supervision matrix $P^i$ being wrongly labeled as 1. Unsurprisingly, this experiment yields a poor instance segmentation performance (a $Pixel$ $F_1$ score




(a) False negative detection for histology images with poor staining

(b) False negative detection for incomplete mitotic cells in histology images

(c) False positive detection due to the difficulty in distinguishing mitotic cells from normal cells

Fig. 12. Due to tissue variability and staining variability, it is difficult for pathologists to collect mitotic cells of all lesion types. Hence, error predictions of our model are inevitable, and we hope to provide some interpretability when our model fails. Error predictions of our model. (a) shows false negative detection for samples containing mitotic cells with poor staining. (b) shows false negative detection for incomplete mitotic cells in histology images. (c) shows false positive detection that incorrectly identifies non-mitotic cells as mitotic cells.

of 0.516 and 11.310% mask AP). If we increase $\tau$ to 0.1, the proportion of adjacent pixel pairs that are wrongly labeled as 1 in $P^i$ Significantly drops. As a result, the model can yield high-quality instance masks, achieving a $Pixel\ F_1$ score of 0.761 and 52.189% mask AP. If we continue to increase $\tau$ to 0.2 or 0.3, we can be more confident that these positive elements in $P^i$ do share the same labels. Our method is not sensitive to the hyper-parameter $\tau$, and a better performance can be obtained when $\tau$ is set to 0.2. Although previous experiments have set $\tau$ to 0.3 by default, the performance is similar.

*2) Input Image Size:* Before inputting an image with 512 × 512 pixels into our model, we up-sample it into 1024 × 1024 pixels. To seek the appropriate scale, we train our model with different scales on TUPAC16 dataset, and results are presented in Table XXI. Higher pixel resolution can help improve the detection performance indeed. Especially for small objects, increasing pixel resolution can ensure that feature maps extracted by the network contain enough information to distinguish background from objects. It can be observed that the performance improves drastically with the scales increasing to 2. Further increasing the scale of input images no longer provides more performance gain.

## V. Instruction for Clinicians

We will discuss the principle and interpretability of our method in this section.

### A. Interpretability When the Model Fails

Most deep learning-based mitosis detection methods require numerous mitotic sample data for supervised learning. However, mitosis datasets annotated by pathologists are severely insufficient. Moreover, due to tissue variability and staining variability, it is difficult for pathologists to collect mitotic cells of all lesion types. Such a dilemma makes it difficult to rule out the presence of abnormal mitotic cells. Hence, error predictions of our model are inevitable, and we hope to provide some interpretability when our model fails. There are two failure types, including false negative and false positive.

TABLE XX
ABLATION STUDY OF THE HYPER-PARAMETER $\tau$ IN THE CONSISTENCY LOSS ON MITOS12 DATASET

| Method | $F_1$-$score$ | Pixel $F_1$ | Mask AP |
|---|---|---|---|
| CondInst [15] | 0.832 | 0.793 | 55.804 |
| $\boldsymbol{\tau = 0}$ | 0.821 | 0.516 | 11.310 |
| $\boldsymbol{\tau = 0.1}$ | 0.826 | 0.761 | 52.189 |
| $\boldsymbol{\tau = 0.2}$ | **0.835** | 0.767 | **52.640** |
| $\boldsymbol{\tau = 0.3}$ | 0.825 | **0.769** | 52.168 |
| $\boldsymbol{\tau = 0.4}$ | 0.821 | 0.761 | 51.152 |

CondInst [15] is a mask-supervised model. Varying the threshold of color similarity $\tau$ from 0.1 to 0.4, our box-supervised model shows robust segmentation performance.

TABLE XXI
ABLATION STUDY OF THE INPUT IMAGE SIZE ON TUPAC16 DATASET

| Input size | 512 | 768 | 1024 | 1280 |
|---|---|---|---|---|
| Scale | 1 | 1.5 | 2 | 2.5 |
| $F_1$-$score$ | 0.740 | 0.757 | **0.782** | **0.791** |

*1) False Negative Detection:* Although H&E staining is a relatively simple staining method,various artifacts can interfere with a good stain. As shown in Fig.12(a), there is a negative impact on mitosis detection for samples with poor staining. Besides, in Fig.12(b), for incomplete mitotic cells in histology images, our model may be confused about it, resulting in missed detections. Thus, feeding overlapping patches with good staining to our model can considerably avoid failing to detect mitotic cells.

*2) False Positive Detection:* As shown in Fig.12(c), due to the large intra-class variability of mitotic cells and the difficulty in distinguishing mitotic cells from normal cells, our model may incorrectly identify non-mitotic cells as mitotic cells. However, even for diagnostic results from different pathologists, it's poorly reproducible due to individual experience and subjective judgment of pathologists. Thus, more lesion types of mitotic cells may further strengthen the ability of mitosis discrimination.

*3) Explore Probable Mitotic Cells:* As a condition diagnosis issue, the missed detection problem generally has a more serious impact. Our method can reduce the omission of mitosis candidates when setting a lower confidence threshold, which is convenient for pathologists to do secondary filtering. Using TUPAC16 dataset as an example, there are 90 mitotic cells in our validation data. When we set the threshold to 0.1, our model predicts 418 candidates with a high recall of 0.978.

### B. A Free Performance Boost on Your Dataset

We have proposed CDAM to improve the cross-domain detection robustness, making our model obtain better performance for external datasets. Moreover, in addition to using the model trained on TUPAC16 dataset directly, clinicians can also choose to fine-tune the trained model on their pathology images. It is worth noting that CDAM allows clinicians to train their datasets through unsupervised training. Clinicians can obtain cross-domain features between TUPAC16 dataset and their own dataset, resulting in a free performance boost.

## VI. CONCLUSION

In this work, we propose a domain adaptive box-supervised instance segmentation network (DBIN) for mitosis detection, which contains several core modules (BIA head, PMS head, and CDAM). This method allows precise pixel-wise prediction of mitosis using weak labels, providing more detailed evidence for downstream analysis and medical diagnosis. Moreover, for the domain migration problem between various pathology labs, we apply domain adaptation for pixel-level feature alignment. The state-of-the-art results of our DBIN across multiple datasets indicate the effectiveness of taking the mitosis detection task as an application of instance segmentation combined with unsupervised domain adaptation.

## REFERENCES

[1] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up," *Histopathology*, vol. 19, no. 5, pp. 403–410, 1991.
[2] M. Aubreville *et al.*, "Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, Dec. 2020.
[3] C. Sommer, L. Fiaschi, F. A. Hamprecht, and D. W. Gerlich, "Learning-based mitotic cell detection in histopathological images," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 2306–2309.
[4] M. Veta, P. J. van Diest, and J. P. Pluim, "Detecting mitotic figures in breast cancer histopathology images," *Proc. SPIE*, vol. 8676, May 2013, Art. no. 867607.
[5] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2013, pp. 411–418.
[6] D. Tellez *et al.*, "Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2126–2136, Sep. 2018.
[7] C. Li, X. Wang, W. Liu, L. J. Latecki, B. Wang, and J. Huang, "Weakly supervised mitosis detection in breast histopathology images using concentric loss," *Med. Image Anal.*, vol. 53, pp. 165–178, 2019.
[8] H. Chen, Q. Dou, X. Wang, J. Qin, and P. A. Heng, "Mitosis detection in breast cancer histology images via deep cascaded networks," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016.
[9] C. Li, X. Wang, W. Liu, and L. J. Latecki, "DeepMitosis: Mitosis detection via deep detection, verification and segmentation networks," *Med. Image Anal.*, vol. 45, pp. 121–133, Apr. 2018.
[10] T. Mahmood, M. Arsalan, M. Owais, M. B. Lee, and K. R. Park, "Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs," *J. Clin. Med.*, vol. 9, no. 3, p. 749, Mar. 2020.
[11] B. E. Bejnordi *et al.*, "Stain specific standardization of whole-slide histopathological images," *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 404–415, Feb. 2016.
[12] M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, and M. Veta, "Domain-adversarial neural networks to address the appearance variability of histopathology images," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 83–91.
[13] X. Zhao, S. Liang, and Y. Wei, "Pseudo mask augmented object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4061–4070.
[14] C.-C. Hsu, K.-J. Hsu, C.-C. Tsai, Y.-Y. Lin, and Y.-Y. Chuang, "Weakly supervised instance segmentation using the bounding box tightness prior," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 6586–6597.
[15] Z. Tian, C. Shen, X. Wang, and H. Chen, "BoxInst: High-performance instance segmentation with box annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5443–5452.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[17] J. Pont-Tuset, P. Arbeláez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, Jan. 2017.

[18] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[19] Y. Zhang *et al.*, "From whole slide imaging to microscopy: Deep microscopy adaptation network for histopathology cancer image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2019, pp. 360–368.

[20] T. Wang, X. Zhang, L. Yuan, and J. Feng, "Few-shot adaptive faster R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7173–7182.

[21] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3339–3348.

[22] C.-C. Hsu, Y.-H. Tsai, Y.-Y. Lin, and M.-H. Yang, "Every pixel matters: Center-aware feature alignment for domain adaptive object detector," in *Proc. SPIEEuropean Conf. Comput. Vis.* Springer, 2020, pp. 733–748.

[23] M. Aubreville, C. A. Bertram, T. A. Donovan, C. Marzahl, A. Maier, and R. Klopfleisch, "A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research," *Sci. Data*, vol. 7, no. 1, pp. 1–10, Dec. 2020.

[24] *ICPR 2014 Mitosis Detection Dataset*. Accessed: Dec. 3, 2021. [Online]. Available: https://mitos-atypia-14.grand-challenge.org/home/

[25] L. Roux *et al.*, "Mitosis detection in breast cancer histological images an ICPR 2012 contest," *J. Pathol. Informat.*, vol. 4, no. 1, p. 8, 2013.

[26] M. Veta *et al.*, "Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge," *Med. Image Anal.*, vol. 54, pp. 111–121, May 2019.

[27] M. Aubreville *et al.*, "Mitosis domain generalization challenge," 2021, Zenodo, Tech. Rep., doi: 10.5281/zenodo.4573978.

[28] X. Li, S. Lai, and X. Qian, "DBCFace: Towards pure convolutional neural network face detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1792–1804, Apr. 2022.

[29] Y. Xue, G. Bigras, J. Hugh, and N. Ray, "Training convolutional neural networks and compressed sensing end-to-end for microscopy cell detection," *IEEE Trans. Med. Imag.*, vol. 38, no. 11, pp. 2632–2641, Nov. 2019.

[30] Z. Yan, X. Yang, and K.-T. Cheng, "Enabling a single deep learning model for accurate gland instance segmentation: A shape-aware adversarial learning framework," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 2176–2189, Jun. 2020.

[31] Y. Su, Y. Lu, J. Liu, M. Chen, and A. Liu, "Spatio-temporal mitosis detection in time-lapse phase-contrast microscopy image sequences: A benchmark," *IEEE Trans. Med. Imag.*, vol. 40, no. 5, pp. 1319–1328, May 2021.

[32] Y. Xue, Y. Li, S. Liu, P. Wang, and X. Qian, "Oriented localization of surgical tools by location encoding," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 4, pp. 1469–1480, Apr. 2022.

[33] A. Sohail, A. Khan, N. Wahab, A. Zameer, and S. Khan, "A multi-phase deep CNN based mitosis detection framework for breast cancer histopathological images," *Sci. Rep.*, vol. 11, no. 1, pp. 1–18, Dec. 2021.

[34] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, Apr. 2016.

[35] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K.: Springer, Aug. 2020, pp. 282–298.

[36] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 20, 2021, doi: 10.1109/TPAMI.2021.3074313.

[37] D. Zhang, J. Han, L. Zhao, and T. Zhao, "From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5549–5560, Dec. 2020.

[38] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 363–380, 2018.

[39] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 876–885.

[40] P. P. Busto and J. Gall, "Open set domain adaptation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 754–763.

[41] X. Fang, H. Bai, Z. Guo, B. Shen, S. Hoi, and Z. Xu, "DART: Domain-adversarial residual-transfer networks for unsupervised cross-domain image classification," *Neural Netw.*, vol. 127, pp. 182–192, Jul. 2020.

[42] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5001–5009.

[43] Y. Zhang *et al.*, "Collaborative unsupervised domain adaptation for medical image diagnosis," *IEEE Trans. Image Process.*, vol. 29, pp. 7834–7844, 2020.

[44] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[45] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9157–9166.

[46] D. Cai, X. Sun, N. Zhou, X. Han, and J. Yao, "Efficient mitosis detection in breast cancer histology images by RCNN," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 919–922.

[47] F. Wilm, C. Marzahl, K. Breininger, and M. Aubreville, "Domain adversarial RetinaNet as a reference algorithm for the MItosis DOmain generalization challenge," 2021, *arXiv:2108.11269*.

[48] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[49] X. Li *et al.*, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," 2020, *arXiv:2006.04388*.

[50] H. Wang *et al.*, "Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection," *Proc. SPIE*, vol. 9041, Mar. 2014, Art. no. 90410B.

[51] A. Paul, A. Dey, D. P. Mukherjee, J. Sivaswamy, and V. Tourani, "Regenerative random forest with automatic feature selection to detect mitosis in histopathological breast cancer images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 94–102.

[52] N. Wahab, A. Khan, and Y. S. Lee, "Transfer learning based deep CNN for segmentation and detection of mitoses in breast cancer histopathological images," *Microscopy*, vol. 68, no. 3, pp. 216–233, Jun. 2019.

[53] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.