IRNet: Instance Relation Network for Overlapping Cervical Cell Segmentation

Yanning Zhou[1], Hao Chen (✉)[2], Jiaqi Xu[1], Qi Dou[3], Pheng-Ann Heng[1,4]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
{ynzhou, hchen}@cse.cuhk.edu.hk
[2]Imsight Medical Technology, Co., Ltd. Hong Kong SAR, China
[3]Department of Computing, Imperial College London, London, UK
[4]Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Abstract. Cell instance segmentation in Pap smear image remains challenging due to the wide existence of occlusion among translucent cytoplasm in cell clumps. Conventional methods heavily rely on accurate nuclei detection results and are easily disturbed by miscellaneous objects. In this paper, we propose a novel Instance Relation Network (IRNet) for robust overlapping cell segmentation by exploring instance relation interaction. Specifically, we propose the Instance Relation Module to construct the cell association matrix for transferring information among individual cell-instance features. With the collaboration of different instances, the augmented features gain benefits from contextual information and improve semantic consistency. Meanwhile, we proposed a sparsity constrained Duplicate Removal Module to eliminate the misalignment between classification and localization accuracy for candidates selection. The largest cervical Pap smear (CPS) dataset with more than 8000 cell annotations in Pap smear image was constructed for comprehensive evaluation. Our method outperforms other methods by a large margin, demonstrating the effectiveness of exploring instance relation.

1 Introduction

Pap smear test is extensively used in cervical cancer screening to assist premalignant and malignant grading [11]. By estimating the shape and morphology structure, e.g., nuclei to cytoplasm ratio, cytologists can give a preliminary diagnosis and facilitate the subsequent treatment. Given that this work is time-consuming and has large intra-/inter-observer variability, designing automatic cell detection and segmentation methods is a promising way towards accurate, objective and efficient diagnosis. However, it remains challenging because the multiple layers of cells partially occlude each other in the Pap smear image, while in H&E image cells do not have multiple layers of translucent overlap. The widely existence of cell clumps along with the translucent cytoplasm raises

obstacles to accurately find the cell boundary. In addition, apart from the target cervical cells, other miscellaneous instances such as white blood cells, mucus and other artifacts are also scattered in the image, which requires an algorithm robust enough to identify them from targets.

Previously, most of the overlapping cell segmentation methods in Pap smear image utilize the shape and intensity information and can be divided into the following steps: cell clump segmentation, nuclei detection and cytoplasm boundary refinement [3,10,13]. However, they demand the precise nuclei detection results as the seeds for the further cytoplasm partition and refinement, which is easily disturbed by the mucus, blood and other miscellaneous instances in clinical diagnosis. Many deep learning based methods have been proposed for gland/nuclei instance segmentation tasks [2,7,12]. Raza et al. proposed Micro-Net for general segmentation task and achieved good results for cell, nuclei and gland segmentation [12]. But it cannot tackle overlapping instances where one pixel could be assigned to multiple instance IDs. On the other hand, the proposal based method can assign multiple labels to a single pixel, which has shown promising results in general object segmentation task. [4] firstly extended the detection method with a segmentation head for instance segmentation, [9] proposed a powerful feature aggregation network backbone. Akram et al. presented the CSPNet consisting of two sub-nets for proposal generation and segmentation nuclei respectively in microscopic image [1]. However, in the Pap smear image, directly extracting in-box features from cell clumps for further processing is not informative enough to distinguish the foreground/background cytoplasm fragment. Meanwhile, the large appearance variance between the single cell and clumps make features semantically inconsistent, which eventually leads to the ambiguous boundary prediction. Besides, it is easy for greedy Non-Maximum Suppression (NMS) to reject true positive predictions in heavy cluster regions due to the misalignment between classification and localization accuracy. Motivated by clinical observation that the appearance of each independent cervical cell in Pap smear image has strong similarity, it shows the potential of leveraging relation information which has been shown effectiveness in other tasks [14,5,6] for better feature representation. For the first time we introduce relation interaction to instance segmentation task and present the Instance Relation Network (IRNet) for overlapping cervical cell segmentation. A novel Instance Relation Module (IRM) is introduced which computes the class-specific instance association for feature refinement. By transferring information among instances using self-attention mechanism, the augmented feature takes merit of contextual information and increase semantic consistency. We also proposed the Duplicate Removal Module (DRM) with the sparsity constraint to benefit proposal selection by calibrating the misalignment between classification score and localization accuracy. To the best of our knowledge, the IRNet is the first end-to-end deep learning method for overlapping cell segmentation in Pap smear image.
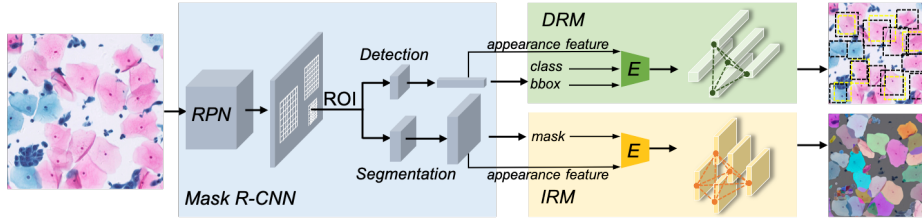
Fig. 1. Overview of the proposed IRNet.

2 Method

As shown in Fig. 1, the proposed IRNet conforms to the two-stage proposal based instance segmentation paradigm [4]. The input image is firstly fed into the Region Proposal Network (RPN) to generate object candidates. Then the candidate features are extracted by the RoIAlign layer [4] and passed through two branches for detection and segmentation. To strengthen the network's ability of candidate selection in cell clumps and improve the semantic consistency among in-box features, we leverage the contextual information among different cells by adding Duplicate Removal Module (DRM) and Instance Relation Module (IRM) after detection and segmentation head.
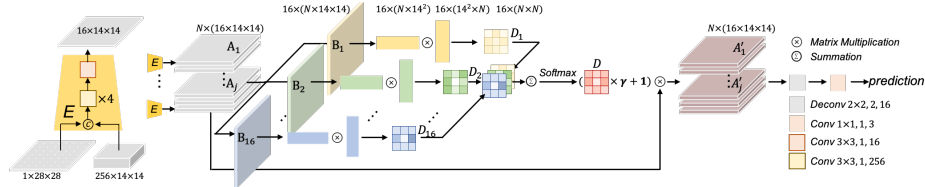
2.1 Instance Relation Module

Fig. 2. Detail structure of the Instance Relation Module in IRNet.

Utilizing in-box features to generate each mask independently is susceptible for cell clumps due to the low foreground contrast and the overlapping boundaries, which eventually leads to ambiguous predictions. Directly enlarging the anchor size to add context won't help a lot in the overlapping region since the surroundings are cells with low contrast. Given that nuclei share the strong appearance similarity so as the cells (shape, texture), we hypothesize that utilizing contextual information from other instance as guidance can increase semantic consistency, especially from those well-differentiated cells. Therefore, we propose the Instance Relation Module (IRM) to exploit the collaborative interaction of

instances. Generally speaking, the IRM takes embedded features from each instance to calculate the instance association matrix, then parses message among features according to their instance relations.

Specifically, the encoder (denotes as **E** in Fig. 2) takes the combination of the predicted mask and deep features as the input to generate the fused features for each candidate. Let n denotes the number of instances in the image, the self-attention mechanism [14]is used to build the association among n instances. As can be seen in Fig. 2, the IRM firstly aggregated the fused features in channel-wise to construct 16 features, denoted as B_j, $j = 1, 2 \ldots 16$, with the shape of $n \times 14 \times 14$. For each B_j, it is reshaped to $\mathbb{R}^{c \times hw}$ and multiplied with its transpose matrix to calculate the channel-wise instance associations, $D_j = B_j B_j^T$. The overall instance association matrix is finally obtained by averaging among all the channel-wise association matrices followed by a Softmax layer for normalization, $D = Softmax(avg(D_1, D_2 \ldots D_c))$. Therefore, the impact of the q-th instance to the p-th instance is computed as $w_{pq} = \frac{exp(d_{pq})}{\sum_k exp(d_{pk})}$, where d_{pq} represents the p-th row, q-th column entry. Let A_p and A_q denotes the p-th and the q-th instance features, A_p' denotes the p-th instance features after relation interaction, the message parsing process can be formulated as:

$$A_p' = \gamma \sum_{q=1}^{n} w_{pq} A_q + A_p, \qquad (1)$$

where γ denotes a learnable scalar factor.

By associating with all the instances, the augmented feature takes merit of contextual information from other instance areas to increase semantic consistency. It is then passed through one 2×2 deconvolution layer with a stride of two, followed by a convolution layer serving as the classifier to output the predicted masks. During training, the Binary Cross-Entropy (BCE) loss is calculated on the ground truth class of masks ($\mathcal{L}_{IRM}$).

2.2 Sparsity Regularized Duplicate Removal Module

Directly utilizing objectness score for NMS leads to sub-optimal results due to the misalignment between classification and localization accuracy, which is more severe for proposals in cell clumps. To calibrate the misalignment, [5] proposed the Duplicate Removal Module (DRM) which takes appearance and location features as input and then utilizes transformed features after relation interaction to directly predict the proposal be *correct* or the *duplicate* using BCE loss ($\mathcal{L}_{DRM}$). The motivation for adding DRM is that the cells and corresponding nuclei have highly correlated spatial distribution.

Based on the observation that cells in Pap smear image gather in several local small clusters instead of one large clump, we propose to add a sparsity constraint on DRM to let the module focus on interaction among the subset of proposals. Specifically, for each target, it only takes proposals with relation weight ranked in the top k for message parsing, where we set $k = 40$ in the experiments. Meanwhile, instead of directly utilizing predicted probability for duplicate removal,

we use the multiplication of classification score and predicted probability for NMS to give a hard constraint of overlapping ratio. Notice that the DRM also exploits the relation information. The proposed IRM is significantly different to adapt to instance segmentation by utilizing a fully convolutional encoder to combine features and predicted masks which effectively preserves the location information and strengthens the effort of shape information.

2.3   Overall Loss Function and Optimization

A multi-task loss is defined as $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{seg} + \alpha\mathcal{L}_{DRM} + \beta\mathcal{L}_{IRM}$, where $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ denote the BCE loss and smooth L1 loss for classification and regression in detection head, and $\mathcal{L}_{seg}$ denotes pixel-wise BCE loss in segmentation head, which are identical as those defined in [4]. $\mathcal{L}_{DRM}$ denotes the BCE loss for *correct* or *duplicate* classification in DRM, where we define the *correct* as the predicted bounding box with the maximum Intersection over Union to the corresponded grounding truth, while others are *duplicate*. L_{IRM} is the pixel-wise BCE loss for refined masks after IRM. α and β are hyper-parameters term for balancing loss weights.

3   Experiments and Results

Dataset and evaluation metrics. The liquid-based Pap test specimen was collected from 82 patients to build the CPS dataset. The specimen was imaged in $\times40$ objective to give the image resolution of around 0.2529 μm per pixel. Then they were cropped into 413 images with the resolution of 1000×1000. In all, 4439 cytoplasm and 4789 nuclei were annotated by the cytologist. To the best of our knowledge, there is no public cervical cell dataset with the annotations on the same order of magnitude to the CPS dataset. To evaluate the proposed method, we split the dataset in patient-level with the ratio of 7:1:2 into the train, valid and test set.

For quantitative evaluation, Average Jaccard Index (AJI) is used which considers in both pixel and object level [7]. AJI$= \frac{\sum_{i=1}^{n} G_i \bigcap S_j}{\sum_{i=1}^{n} G_i \bigcup S_j + \sum_{k \in N} S_k}$, where G_i is the i-th ground truth, S_j is the j-th prediction, $j = \text{argmax}_k \frac{G_i \bigcap S_k}{G_i \bigcup S_k}$. It measures the ratio of the aggregated intersection and aggregated union for all the predictions and ground truths in the image. F1-score (F1) is used to measure the detection accuracy for reference [2]. **Implementation details.** We implemented the proposed IRNet with PyTorch 1.0. The network architecture is the same as [8] in the condition of Feature Pyramid Network with 50-layer ResNet (ResNet-50-FPN). One NVIDIA TITIAN Xp graphic card with CUDA 9.0 and cuDNN 6.0 was used for the computation. During training, we used SGD with 0.9 momentum as the optimizer. The initial learning rate was set as 0.0025 with a factor of 2 for the bias, while the weight decay was set 0.0001. We linear warmed up the learning rate in the first 500 iterations with a warm-up factor of $\frac{1}{3}$.
Effectiveness of the proposed IRNet. Firstly, we conducted experiments to compare different algorithms for overlapping cell segmentation. (1).*JOMLS* [10]:

Table 1. Quantitative comparison against other methods on the test set.

Method	AJI		F1	
	Cyto	Nuclei	Cyto	Nuclei
JOMLS [10]	0.1974	0.3167	0.3794	0.3618
CSPNet [1]	0.4607	0.3891	0.5307	0.5942
Mask R-CNN [4]	0.6845	0.5169	0.6664	0.7192
IRNet w/o IRM	0.6887	0.5342	0.7266	0.7424
IRNet w/o DRM	0.6995	0.5471	0.7010	0.7501
IRNet	**0.7185**	**0.5496**	**0.7497**	**0.7554**

an improved joint optimization of multiple level set functions for the segmentation of cytoplasm and nuclei from clumps of overlapping cervical cells. (2). *CSP-Net* [1]: a cell segmentation proposal network with two CNNs for cell proposal prediction and cell mask segmentation respectively. To give a fair comparison, we reproduce the ResNet-50-FPN instead of the original 6-layer CNN for candidates prediction in the first stage. (3). *Mask R-CNN* [4]: IRNet without Duplicate Removal Module and Instance Relation Module, which can be considered as a standard mask-rcnn structure with ResNet-50-FPN as the backbone. (4). *IRNet w/o IRM*: IRNet without Instance Relation Module. (5). *IRNet w/o DRM*: IRNet without Duplicate Removal Module. (6). *IRNet*: The proposed IRNet.

As can be seen in Table 1, all the deep learning based methods achieve better performance compared with the level-set based method [10]. The reason is that our dataset has more complicated background information including white blood cells and other miscellaneous instances compared with that used in [10], which requires the algorithm to be more robust for the noise. Apart from CSPNet [1] that uses a separate CNN to extract multi-level features in ROI-pooling, our baseline model extracts features from the shared FPN backbone in specific resolution according to the box size, which effectively improves cytoplasm AJI from 0.4607 to 0.6845 and nuclei AJI from 0.3891 to 0.5169. In addition, adding the DRM (*IRNet w/o IRM*) gives a striking 9.03% and 3.23% improvement of F1 score for cytoplasm and nuclei, which proves that the duplicated boxes can be effectively suppressed by parsing message among boxes, especially for the cytoplasm in cell clumps. Compared with baseline model (*Mask R-CNN*), adding IRM after segmentation branch (*IRNet w/o DRM*) gains 2.19% and 5.84% AJI improvements for cytoplasm and nuclei, demonstrates that by leveraging instance association for transferring information, the augmented features are more consistent and discriminative for semantic representation. By combining DRM and IRM in our framework, the proposed IRNet (*Ours*) outperforms other methods by a large margin, with 4.97% and 6.33% improvements of AJI as well as 12.5% and 5.03% improvements of F1 score for cytoplasm and nuclei comparing with the baseline model (*Mask R-CNN*).

Ablation study for the Instance Relation Module. We then conduct the ablation study for the design of the proposed IRM. (1). *DF* (Deep Feature): Using deep features before the deconvolution layer in segmentation branch with

DF	MSK	RL	AJI		F1	
			Cyto	Nuclei	Cyto	Nuclei
			0.6887	0.5342	0.7266	0.7424
✓	✓		0.7042	0.5283	0.7355	0.7306
✓		✓	0.7097	0.5367	0.7350	0.7442
	✓	✓	0.7071	0.5299	0.7324	0.7366
✓	✓	✓	**0.7185**	**0.5496**	**0.7497**	**0.7554**

Table 2. Ablation study for the Instance Relation Module.

the shape of $256 \times 14 \times 14$ for further process. (2). *MSK* (Mask): Using predicted mask from segmentation branch for further process. We keep the masks from the predicted class in detection branch only and remove the other class. (3). *RL* (Relation Learning): Conducting channel-wise instance relation learning. Results are shown in Table 2.

We first add the same number of convolution layers as that in IRM encoder and remove the relation interaction part (*DF + MSK*). Adding more parameters to build deeper layers do improve the results, but all the methods with relation learning have better performance. Compared with *IRNet w/o IRM*, directly utilizing deep features for relation interaction (*DF+RL*) yields the results of 0.7097 and 0.5367 in AJI, which brings 3.05% and 0.47% improvement for cytoplasm and nuclei. Meanwhile, directly using predicted masks (*MSK+RL*) outperforms *IRNet w/o IRM* by 2.67% AJI for cytoplasm. The nuclei class does not improve in *MSK+RL* because the shape of masks are almost the same so that it does not contain enough information for message parsing. Furthermore, when we adopt deep features with selected masks simultaneously, it significantly improves the performance over the (*DF+RL*) by 4.33% and 2.88% for cytoplasm and nuclei, which shows the effectiveness of the IRM design.

Qualitative comparison. Fig. 3 shows representative samples in the test set with challenging cases such as the heavily occlusion of cytoplasm and the scatted white blood cells. Conventional method (*JOMLS*) fails on identifying the miscellaneous instance and the nucleus (see the third row and fourth row). Compared with *Mask R-CNN*, adding IRM mitigates the ambiguous cytoplasm boundary prediction in cell clumps (see (a) and (e)), which is important for further accurate cell classification by morphology analysis. Moreover, combining DRM further suppresses the duplicated predictions successfully (see (e) and (f)).

4   Conclusion

Accurately segmenting the cytoplasm and nucleus instance in Pap smear image is pivotal for further cell morphology analysis and cervical cancer grading. In this paper, we proposed a novel IRNet which leverages the instance relation information for better semantic consistent feature representation. By aggregating the features from other instances,they tend to generate masks with more consistent
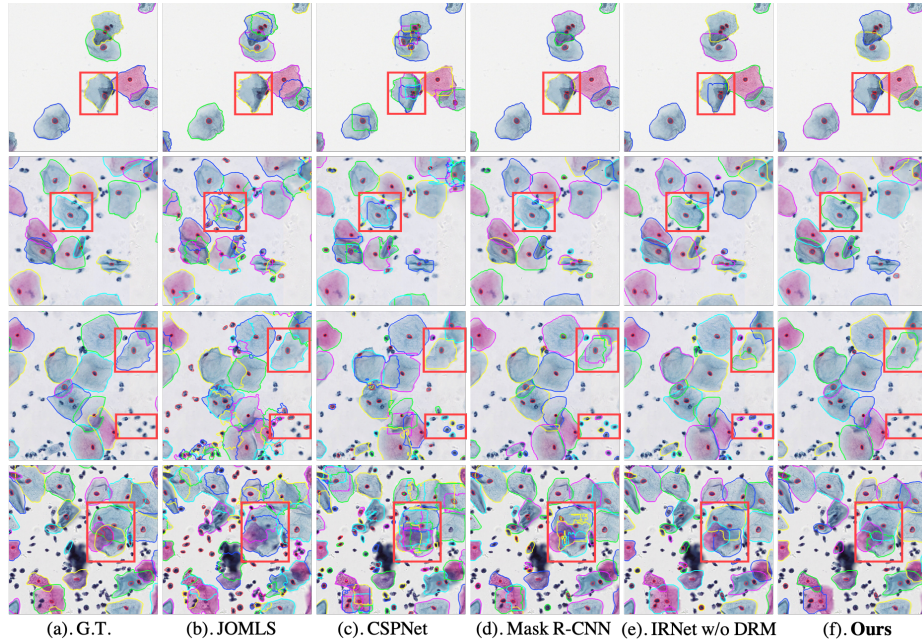
Fig. 3. Qualitative results of Overlapping cervical cell segmentation in Pap smear image on the test set (each closed curve denotes an individual instance). Rectangles show the main differences among different methods.

boundary shape. Quantitative and qualitative results demonstrate the effectiveness of our method. Notice the proposed IRM is inherently general and can be complementary for various proposal-based instance segmentation methods.

5   Acknowledgement

References

1. Akram, S.U., Kannala, J., Eklund, L., Heikkilä, J.: Cell segmentation proposal network for microscopy image analysis. In: Deep Learning and Data Labeling for Medical Applications, pp. 21–29. Springer (2016)

2. Chen, H., Qi, X., Yu, L., Dou, Q., Qin, J., Heng, P.A.: Dcan: Deep contour-aware networks for object instance segmentation from histology images. Medical image analysis **36**, 135–146 (2017)
3. GençTav, A., Aksoy, S., ÖNder, S.: Unsupervised segmentation and classification of cervical cell images. Pattern recognition **45**(12), 4151–4168 (2012)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: IEEE CVPR. pp. 2961–2969 (2017)
5. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: IEEE CVPR. pp. 3588–3597 (2018)
6. Jun Fu, Jing Liu, H.T.Y.L.Y.B.Z.F.a.H.L.: Dual attention network for scene segmentation (2019)
7. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging **36**(7), 1550–1560 (2017)
8. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE CVPR. pp. 2117–2125 (2017)
9. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: IEEE CVPR. pp. 8759–8768 (2018)
10. Lu, Z., Carneiro, G., Bradley, A.P.: An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. IEEE Trans. Image Proc. **24**(4), 1261–1272 (2015)
11. Papanicolaou, G.N.: A new procedure for staining vaginal smears. Science **95**(2469), 438–439 (1942)
12. Raza, S.E.A., Cheung, L., Shaban, M., Graham, S., Epstein, D., Pelengaris, S., Khan, M., Rajpoot, N.M.: Micro-Net: A unified model for segmentation of various objects in microscopy images. Medical Image Analysis **52**, 160–173 (2019)
13. Song, Y., Tan, E.L., Jiang, X., Cheng, J.Z., Ni, D., Chen, S., Lei, B., Wang, T.: Accurate cervical cell segmentation from overlapping clumps in pap smear images. IEEE Trans. Med. Imaging **36**(1), 288–300 (2017)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NIPS. Curran Associates, Inc. (2017)