PCTrans: Position-Guided Transformer with Query Contrast for Biological Instance Segmentation

Qi Chen[1] Wei Huang[1] Xiaoyu Liu[1] Jiacheng Li[1] Zhiwei Xiong[1,2,*]
[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

Abstract

Recently, query-based transformer gradually draws attention in segmentation tasks due to its powerful ability. Compared to instance segmentation in natural images, biological instance segmentation is more challenging due to high texture similarity, crowded objects and limited annotations. Therefore, it remains a pending issue to extract meaningful queries to model biological instances. In this paper, we analyze the problem when queries meet biological images and propose a novel ***P****osition-guided* ***Trans****former with query* ***C****ontrast (****PCTrans****) for biological instance segmentation. PCTrans tackles the mentioned issue in two ways. First, for high texture similarity and crowded objects, we incorporate position information to guide query learning and mask prediction. This involves considering position similarity when learning queries and designing a dynamic mask head that takes instance position into account. Second, to learn more discriminative representation of the queries under limited annotated data, we further design two contrastive losses, namely Query Embedding Contrastive (QEC) loss and Mask Candidate Contrastive (MCC) loss. Experiments on two representative biological instance segmentation datasets demonstrate the superiority of PCTrans over existing methods. Code is available at* https://github.com/qic999/PCTrans.

1. Introduction

Biological instance segmentation is a prerequisite for analyzing the behaviors and properties of target organisms [8, 42, 24]. Compared to instance segmentation in natural images, this task is more challenging due to the variety of uneven texture, ambiguous boundary, and morphological differences. Besides, overlapping and occlusions of instances are severe in different biological image modalities, such as plant phenotype images [38], fluorescence microscope (FM) images [33], Haematoxylin and Eosin (H&E) stained histology slides [44], and electron microscope (EM) images [4, 30]. Therefore, it is highly desirable to design accurate and reliable instance segmentation algorithms for biology and biomedical research.

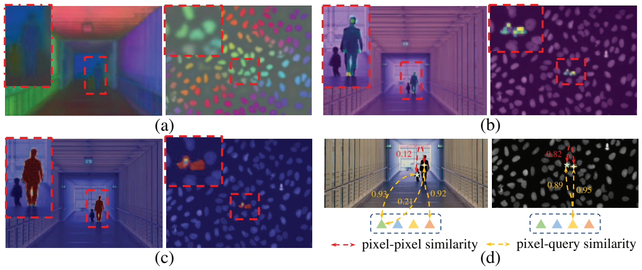

Figure 1. Illustration of our motivation. (a) shows the discriminative difference of pixel embeddings. (b) shows the attention weight of the last cross-attention layer. (c) shows the activation map of the output layer. (d) shows the segmentation error caused by the high similarity between one query and multiple instances feature in biological instance segmentation.

Deep learning-based methods are widely used in biological instance segmentation nowadays. They can be grouped into two categories: proposal-based and proposal-free methods. Limited by bounding boxes, proposal-based methods typically suppress valid objects when adjacent instances have severe overlap. Besides, these methods introduce multiple hyper-parameters and design choices, which are non-trivial to select for different biological datasets [38, 33, 44]. Proposal-free methods get rid of the limitation of bounding boxes and are less sensitive to different object sizes due to post-processing. However, the final segmentation performance is sensitive to the selection of hyper-parameters in post-processing algorithms [12, 2, 48, 5, 18]. Meanwhile, independent post-processing algorithms prevent end-to-end training of the segmentation pipeline, leading to a sub-optimal result. In a word, both types of methods suffer from crowded objects to varying degrees, merging adjacent objects or suppressing valid instances.

Recently, query-based transformer networks present a new paradigm of instance segmentation which aggregates

[*] Corresponding author: zwxiong@ustc.edu.cn.

object-related information and provide a group of object queries to output the final set of mask candidates [7, 6, 37]. This paradigm simplifies the segmentation pipeline and shows excellent results in natural images. The mask candidates are obtained by computing the similarity between pixel embeddings and query embeddings. However, there is a large gap when applying these works to model biological instances features by queries due to high feature similarity between biological instances. As shown in Figure 1 (a), the pixel embeddings in natural images are more discriminative than those in biological images. Therefore, the distinct instance feature can be better aggregated into queries through a cross-attention mechanism (Figure 1 (b)). Furthermore, mask candidates with high quality can be produced due to the more precise activation (Figure 1 (c)). In biological images, queries are easier to aggregate vague instances features with similar pixel embeddings, leading to segmentation error (Figure 1 (d)). In other words, the existing methods more rely on visual similarity in query learning, which is not enough for biological instances.

To tackle the above issue, we propose a novel one-stage position-guided transformer with query contrast (PCTrans) for biological instance segmentation. PCTrans is based on two key designs, including 1) exploring position information to guide query learning and mask predicting, and 2) making queries more discriminative by contrastive learning. To obtain the position attribute of queries, we iteratively predict the center of instances based on queries. For query learning, we consider the position similarity between query position and pixel position with a cross-attention mechanism. In this way, query learning alleviates the over-reliance on visual similarity. For mask predicting, we design a position-aware dynamic mask head conditioned on queries, which can better fuse features between queries and pixel embeddings under the position guidance, adaptively producing high-quality mask candidates. Benefiting from the strong capacity of dynamic conditional convolutions, the mask head can be very lightweight, which is computationally friendly.

Another challenge in using transformers for biological tasks is the limited availability of annotation data. Transformers require a substantial amount of training data to effectively learn discriminative queries and perform well on dense prediction tasks [13, 50, 45]. However, there is currently a scarcity of annotated training data for the biological instance segmentation task. To make up for the disadvantage in data, we design two contrastive losses, namely Query Embedding Contrastive (QEC) loss and Mask Candidate Contrastive (MCC) loss to learn more discriminative queries. The construction of positive and negative query pairs is achieved by a clustering algorithm, regarding the queries best-matched with ground truth as clustering centers. Then we perform contrastive learning to the queries and the corresponding predicted masks, respectively.

Overall, the contributions of this work are summarized as follows:

- We propose the first one-stage position-guided transformer with query contrast (PCTrans) for biological instance segmentation.
- To tackle high texture similarity and crowded objects, we explicitly explore position information of instances to guide query learning and mask predicting.
- To overcome limited annotations, we design two contrastive losses, QEC loss and MCC loss, to enhance queries representation.
- Our proposed PCTrans achieves state-of-the-art performance on two representative biological instance segmentation benchmarks.

2. Related Work

2.1. Biological Instance Segmentation

Previous works on biological instance segmentation mainly follow two directions: proposal-based and proposal-free methods. Proposal-based methods [27, 26, 53, 54, 56] first locate objects by bounding boxes and subsequently refine the instance mask within the region of interest. Proposal-free methods predict well-designed instance-aware features and morphological properties, followed by post-processing algorithms to yield final results. These methods focus on designing elegant networks to obtain high-quality intermediate of instances, such as affinity, gradient map, and so on. With the assistance of complex and well-designed post-processing algorithms such as waterz [12], LMC [2], Mutex [48], and conditional random field [23], these intermediates can be processed into instance masks. The proposal-free methods [9, 22, 29, 49, 31] are prevalent in nuclei segmentation, leaf and plant segmentation, which both are common instance segmentation tasks in biological images. In this paper, we are the first to utilize queries for modeling biological instance features and investigate their potential in biological instance segmentation.

2.2. Query-based Transformer

Transformer [43] were born out of natural language processing and have been successfully extended to the field of computer vision [11]. Recently, DETR [3] is proposed to combine transformer with a CNN backbone to aggregate object-related information and provided a group of object queries to output the final set of predictions in object detection task. MaskFormer [7] introduces DETR structure into the segmentation task and tries to solve the segmentation task in a unified framework. OSFormer [37] introduces DETR structure into camouflaged instance segmentation. In this work, we build upon the idea of the query-based trans-

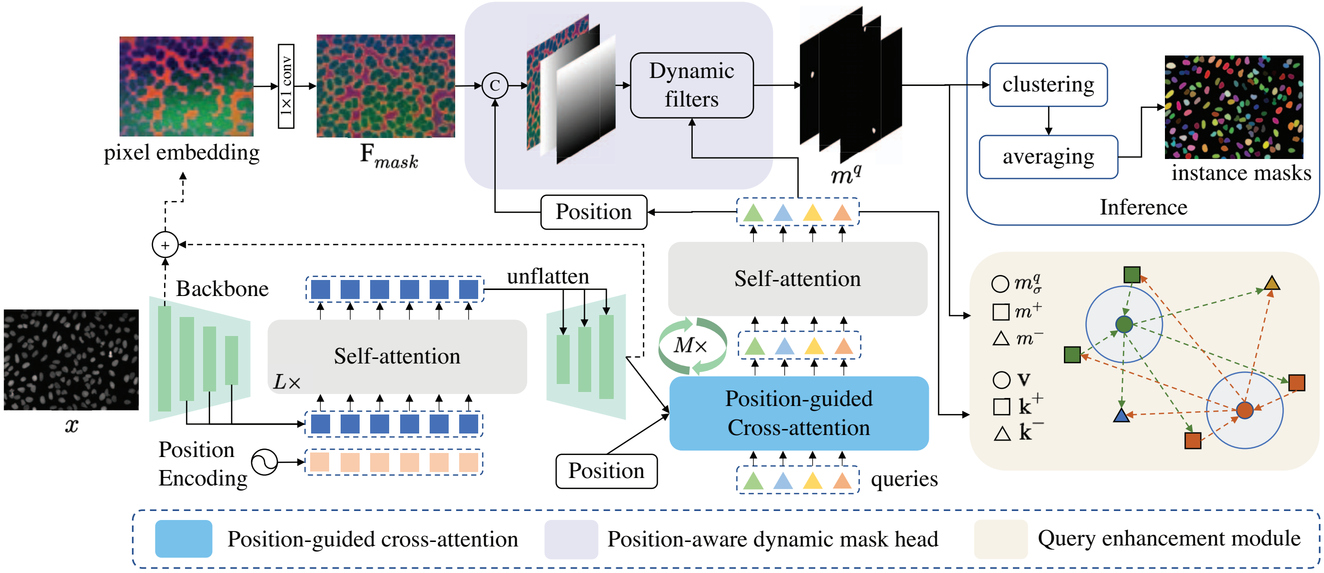

Figure 2. The overall architecture of PCTrans. First, the input image is fed into the backbone and a stack of L self-attention layers to produce multi-scale pixel embeddings with strong feature representation. Then, queries iteratively aggregate distinct instance features from multi-scale pixel embeddings through the position-guided cross-attention layer and interact with each other through the self-attention mechanism for M times. During each iteration, the queries are used to predict the position and masks of instances. At the end of the iteration, the queries and the corresponding predicted masks are regularized by the query enhancement module. At inference time, we adopt a clustering-then-averaging strategy to get the final segmentation results.

former and improve feature extraction and utilization for biological instance segmentation.

2.3. Contrastive Learning

Contrastive learning has shown excellent prospects in representation learning. As a representative, MOCO [14] uses contrastive learning for image-level self-supervised training. [46] and [17] introduce pixel-level and region-aware contrastive learning for semantic segmentation, respectively. However, it does not draw much attention to biological instance segmentation. Inspired by these works, we perform contrastive learning to the instance-aware queries and the corresponding predicted masks, respectively, which can help the network learn more discriminative query features and make up for the disadvantage of the small amount of data in biological instance segmentation.

2.4. Dynamic Convolution

Dynamic filters have been explored in dynamic filter networks [20] and CondConv [52] for classification task. It differs from traditional convolution in that another network dynamically generates the filter. SOLOv2 [47] and CondInst [41] extend this idea to solve the challenges of instance segmentation in neural images which adds an extra kernel branch with the same architecture as the classification head. In this work, to exploit the information from the queries, we introduce dynamic filters conditioned by queries to be aware of position information and produce high-quality mask candidates adaptively. In this way, we can get better prediction masks, further improving the segmentation performance.

3. Proposed Method

In this section, we first formulate the biological instance segmentation task in a query-based framework and present the overall architecture of the proposed PCTrans. Then we describe the details of the position-guided cross-attention mechanism, position-aware dynamic mask head, and query enhancement module of PCTrans. Finally, the training and inference procedure are discussed.

3.1. Problem Formulation

We denote an image-label pair with $\{x, m^{gt}\}$, where x is the input image, $m^{gt}$is the corresponding set of ground truth instance masks. Specifically, m^{gt} can be formulated as $\{m_i^{gt} | m_i^{gt} \in \{0,1\}^{H\times W}\}_{i=1}^{K}$, where H and W represent the height and width of the input image x, and K is the number of instances in the input image x. For every input image x, there will be N queries to represent the K instances. When the image x is fed into the network, we can obtain the mask candidates $m^q = \{m_j^q | m_j^q \in \{0,1\}^{H\times W}\}_{j=1}^{N}$. To train this framework, matching between the set of predictions m^q and the set of ground truth segments m^{gt} is required. Following the common practice in query-based frameworks, we adopt a bipartite matching method [40].

Since there are only foreground and background in the biological instance segmentation task, we directly use mask candidates to compute the assignment costs for the matching problem, *i.e.*, $\text{BCE}(m_i^{gt}, m_j^q) + \text{DICE}(m_i^{gt}, m_j^q)$, where BCE$(\cdot)$ is the binary cross-entropy loss, and DICE$(\cdot)$ is the dice loss [35]. So the bipartite matching-based assignment σ between the set of predictions m^q and m^{gt} is conducted for computing mask loss:

$$\mathcal{L}_m = \sum_{i=1}^{K} (\text{BCE}(m_{\sigma(i)}^q, m_i^{gt}) + \text{DICE}(m_{\sigma(i)}^q, m_i^{gt}). \quad (1)$$

As shown in Figure 2, we keep the process of extracting multi-scale pixel embeddings $\mathbf{X} = \{\mathbf{x}_k\}_{k=1}^{hw}$ through backbone and self-attention following previous works [6, 37], where h and w represent the height and width of the feature map. The difference is that the position-guided cross-attention mechanism and the position-aware dynamic mask head are introduced to explicitly explore position information of instances to guide query learning and mask predicting. Besides, the query enhancement module contains two contrastive losses to make queries more discriminative.

3.2. Position Guidance

To construct position guidance, we first need to obtain the position information of instances. To achieve this target, we intuitively predict the center of instances based on the queries. Inspired by [3], we predict the offset from the reference point to the target center point. The coordinate is predicted from each query as follows,

$$\mathbf{c}_j = \text{sigmoid}(\text{MLP}(\mathbf{f}_j) + \mathbf{s}_j), \quad (2)$$

where $\mathbf{f}_j$ is the updated query. $\mathbf{c}_j$ is a two-dimensional vector $[c_{jx}, c_{jy}]^\top$ and presents the predicted coordinate of the center point. Sigmoid function aims to normalize the prediction $\mathbf{c}_j$ to the range [0, 1]. An MLP is used to predict the unnormalized center point coordinate. $\mathbf{s}_j$ is the unnormalized 2D coordinate of the reference point. In our method, $\mathbf{s}_j$ is the predicted center point coordinate in the last iteration. We define the set of ground truth coordinates as $\mathbf{c}^{gt} = \left\{\mathbf{c}_i^{gt} | \mathbf{c}_i^{gt} \in \{0, 1\}\right\}_{i=1}^{K}$. So we can compute the point loss $\mathcal{L}_p$ as follows:

$$\mathcal{L}_p = \sum_{i=1}^{K} ||\mathbf{c}_{\sigma(i)}^q, \mathbf{c}_i^{gt}||_1. \quad (3)$$

3.2.1 Position-Guided Cross-Attention

To consider position similarity between query position and pixel position in the cross-attention mechanism, we need to compute a position embedding that can represent the current query location attribute. Motivated by [34], we hypothesize this position embedding can be produced by queries and reference point. The insights behind it are that using the sinusoidal positional encoding function can map the reference point to the same embedding space with the positional embedding of pixel embedding. However, only relying on the default embedding of the reference point is not enough to represent the position attribute of the query. So we need to transform the default embedding of the reference point conditioned on the query, and the transformed position embedding can better represent the current query location attribute. The transformed position embedding $\mathbf{f}_j^p$ is computed as follows,

$$\begin{aligned} \mathbf{f}_j^p =& \mathbf{T}_j * \text{sinusoidal}(\text{sigmoid}(\mathbf{s}_j)) \\ =& \text{MLP}(\mathbf{f}_j) * \text{sinusoidal}(\text{sigmoid}(\mathbf{s}_j)). \end{aligned} \quad (4)$$

In this way, we form the query $\tilde{\mathbf{f}}_j$ in position-guided cross-attention mechanism by concatenating the default query $\mathbf{f}_j$ and the transformed position embedding $\mathbf{f}_j^p$.

3.2.2 Position-Aware Dynamic Mask Head

In addition to adding position guidance to query learning, we also consider introducing position information to the process of mask predicting. Inspired by CondInst [41], the dynamic filters controlled by distinct instance features are position-aware, benefiting from their strong nonlinear property. Hence, we introduce dynamic filters to map the pixel embeddings to the mask candidates, which are conditioned on the queries. We adopt compact dynamic filters as the mask head on the given feature map $\mathbf{F}_{mask}$, which only contain a three-layer 1×1 convolution. This compact mask head can provide better nonlinear properties than simple computing similarity between query and pixel embedding to predict masks, resulting in better prediction. The parameters of the mask head are adaptively generated by a controller head which is a simple MLP conditioned by queries. In order to reduce the number of the generated parameters, we get $\mathbf{F}_{mask}$ by reducing the channel number of pixel embedding to C_{mask}. To make full use of the position information, $\mathbf{F}_{mask}$ is combined with a relative coordinate map $\mathbf{M}_j^{rc}$ of the predicted coordinate $\mathbf{c}_j$. Then, the combination is sent to the mask head to predict the instance mask:

$$\begin{aligned} m_j^q =& \text{MaskHead}(\text{Concat}(\mathbf{F}_{mask}, \mathbf{M}_j^{rc}), w_j) \\ =& \text{MaskHead}(\text{Concat}(\mathbf{F}_{mask}, \mathbf{M}_j^{rc}), \text{MLP}(\mathbf{f}_j)), \end{aligned} \quad (5)$$

where w_j is the parameters of the mask head, MLP is the controller head to predict w_i, and Concat $(\cdot, \cdot)$ is the concatenate operator.

3.3. Query Enhancement Module

More discriminative feature representation can facilitate distinguishing instances especially those with overlapping, thereby improving the segmentation performance. To this end, we introduce contrastive learning between queries to make the matched queries belonging to the same object instance closer in embedding space and the unmatched queries farther away. Similarly, we also perform contrastive learning between the prediction masks. Specifically, we propose a Query Embedding Contrastive (QEC) loss for queries and a Mask Candidate Contrastive (MCC) loss for prediction masks to achieve contrastive property.

We observe that if the queries predict the same object, the cosine similarity between the corresponding query is closer to 1. Besides, each instance can be predicted evenly by queries and we can easily use the cosine similarity of the query to cluster the query. We define the query matched with ground truth mask as contrastive embedding $\mathbf{v}$ according to the results of bipartite matching. By comparing the cosine similarity between queries, we cluster the queries into groups regarding the contrastive embedding $\mathbf{v}$ as the clustering center. Queries belonging to the same group as contrastive embedding $\mathbf{v}$ are considered positive samples of equivalent contrastive embedding. The QEC loss for a positive pair of examples is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{QEC}} &= -\sum_{\mathbf{k}^+} \log \frac{e^{d(\mathbf{v},\mathbf{k}^+)/\tau_q}}{e^{d(\mathbf{v},\mathbf{k}^+)/\tau_q} + \sum_{\mathbf{k}^-} e^{d(\mathbf{v},\mathbf{k}^-)/\tau_q}} \\ &= -\log\left[1 + \sum_{\mathbf{k}^+}\sum_{\mathbf{k}^-} e^{d\left(\mathbf{v},\mathbf{k}^-\right)/\tau_q - d\left(\mathbf{v},\mathbf{k}^+\right)/\tau_q}\right],\end{aligned} \tag{6}$$

where τ_q is a temperature hyper-parameter to control the scale of terms, $\mathbf{k}^+$ and $\mathbf{k}^-$ are positive and negative queries, respectively, and $d(\mathbf{v},\mathbf{k}^+)$ denotes the cosine similarity distance.

Similarly, the MCC loss for multiple positive examples is defined as follows:

$$\mathcal{L}_{\text{MCC}} = -\log\left[1 + \sum_{m_i^+}\sum_{m_i^-} e^{d(m^q_{\sigma(i)},m_i^-)/\tau_m - d(m^q_{\sigma(i)},m_i^+)/\tau_m}\right], \tag{7}$$

where τ_m is a temperature hyper-parameter to control the scale of terms, m_i^+ and m_i^- are positive and negative query masks, respectively, and $d(m^q_{\sigma(i)}, m_i^-) = \frac{2\left|m^q_{\sigma(i)} \cap m_i^-\right|}{\left|m^q_{\sigma(i)}\right| \cup \left|m_i^-\right|}$ measures the similarity between mask candidates.

3.4. Training and Inference

To make the network produce better pixel embedding, we add two auxiliary losses, including discriminative loss [9] $\mathcal{L}_d$ and semantic loss $\mathcal{L}_s$ on the pixel embedding. Finally, the whole model is optimized with a multi-task loss function:

$$\mathcal{L} = \lambda_1\mathcal{L}_m + \lambda_2\mathcal{L}_p + \lambda_3\mathcal{L}_d + \lambda_4\mathcal{L}_s + \lambda_5\mathcal{L}_{\text{QEC}} + \lambda_6\mathcal{L}_{\text{MCC}}, \tag{8}$$

where λ_{1-6} are weighting coefficients to balance these six terms.

In the inference phase, there is no ground truth mask and we cannot know which query mask is the best candidate for the objects. To maximize the performance advantage of our network, we propose a simple but effective strategy to exploit the total query mask. The process of multiple queries predicting the same object is like the procedure of test augmentation in semantic segmentation. Therefore, we can cluster queries in a simple threshold-based clustering way to get all queries that predict the same instance. Then we only need to average the masks in the same category to get the final instance masks.

4. Experiments

4.1. Dataset and Metric

Fluorescence Microscopy Images. BBBC039V1 [33] is part of a high-throughput chemical screen on U2OS cells, with examples of 200 bioactive compounds. The effect of the treatments was originally imaged using the Cell Painting assay (fluorescence microscopy). This dataset contains 200 images in size 520×696 which present a variety of nuclear phenotypes, representative of high-throughput chemical perturbations. Following [19], we use 100 images for training, 50 images for validation, and the rest of the 50 images for testing. Following the existing methods, we adopt four common metrics, including Aggregated Jaccard Index (AJI), pixel-level Dice score (Dice), object-level F1 score (F1) and Panoptic Quality (PQ).

Plant Phenotype Images. The CVPPP A1 dataset [38] is one of the most common instance segmentation benchmarks, which consists of 128 training images and 33 testing images with a size of 530×500 pixels. Following [19], we randomly select 20 images from the training set as the validation set. This dataset is challenging due to the high variety of leaf shapes and severe occlusion among leaves. The quality of the segmentation result is measured by Symmetric Best Dice (SBD) and absolute Difference in Counting (DiC) metrics.

4.2. Implementation Details

For multi-scale pixel embedding extraction, we use multi-scale deformable attention Transformer (MSDeformAttn) [57] as the self-attention layer. Specifically, we use 6 MSDeformAttn layers applied to feature maps with resolution 1/8, 1/16 and 1/32 from the Backbone. Besides, we use a simple downsampling layer with a lateral connection on the 1/4 feature map to generate the feature map of resolution 1/8 as the pixel embedding. To get the feature map $\mathbf{F}_{mask}$, we adopt a 1×1 convolution to reduce the channel of pixel embedding to 16. In addition, we set the channels of the dynamic filters all as 8 by default. We use 9 position-guided cross-attention layers and standard self-attention layers for querying learning. Besides, we set

Method	AJI ↑	Dice ↑	F1 ↑	**PQ** ↑
Mask RCNN[15]	0.7983	0.9277	0.9180	0.7773
Cell RCNN[55]	0.8070	0.9290	0.9276	0.7959
UPSNetN[51]	0.8128	0.9274	0.9191	0.7857
JSISNet[10]	0.8134	0.9316	0.9282	0.7913
PanFPN[21]	0.8193	0.9320	0.9275	0.7960
OANet[28]	0.8198	0.9372	0.9330	0.8085
AUNet[25]	0.8252	0.9377	0.9315	0.8090
Cell RCNNv2[27]	0.8260	0.9336	0.9328	0.8010
PFFNet[26]	0.8477	0.9478	0.9451	0.8331
PEA [19]	0.8674	0.9473	-	0.8420
BISSG [32]	0.8680	0.9482	**0.9670**	0.8629
OSFormer [37]	0.7414	0.9206	0.8755	0.7516
Mask2Former [6]	0.7671	0.9601	0.8879	0.8011
Ours	**0.9022**	**0.9625**	0.9668	**0.8922**

Table 1. Quantitative comparison with state-of-the-art methods on the test set of BBBC039V1.

300 queries and 100 queries on BBBC039V1 and CVPPP datasets, respectively. For all experiments, we set the batch size to 8. We adopt the AdamW [**?**] optimizer and the step learning rate schedule. We use an initial learning rate of 0.0001 with linear warming up in the first 1000 iterations and a weight decay of 0.05 for all layers. We adopt a crop size of 512 × 512 and train 30k iterations on BBBC039V1. We use a crop size of 448 × 448 and train 60k iterations on CVPPP. Following [19], we adopt the same data augmentation for BBBC039V1 and CVPPP. Following [6], we calculate the mask loss with sampled points in both the matching and the final loss calculation to save GPU memory and improve training efficiency. We set $\lambda_1 = \lambda_2 = \lambda_4 = 5$ and $\lambda_3 = \lambda_5 = \lambda_6 = 2$ by default. When compared with state-of-the-art methods, we set $\tau_q = 2, \tau_m = 0.5$ and adopt ResNet-101 [16] as the backbone. In the ablation study, we set $\tau_q = \tau_m = 1$ and adopt ResNet-50 [16] as the backbone.

4.3. Comparison with State-of-the-art Methods

4.3.1 Comparison with CNN-based Methoeds

Results on BBBC039V1. We demonstrate the effectiveness of our method on the BBBC039V1 dataset. As shown in Table 1, our proposed PCTrans achieves the best performance on all metrics. The performance of PCTrans is obviously improved compared with the latest proposal-based method PFFNet [26] and the proposal-free methods PEA [19] and BISSG [32]. Specifically, PCTrans improves the key PQ metric by 2.48%. We further carry out a qualitative visualization analysis. As shown in the first example in Figure 3, PCTrans achieves better results for nuclear instance segmentation compared to these three latest methods. Specifically, compared to PEA and BISSG, our PCTrans effectively distinguishes nuclear pixels from the background and segments different instances especially those with overlapping. Compared to the proposal-based method PFFNet, our PCTrans can segment more precise instance contours.

Method	**SBD** ↑	|DiC| ↓
Nottingham[39]	68.3	3.8
IPK[36]	74.4	2.6
AC[1]	79.1	1.1
Discriminative* [9]	79.6	1.4
PEA* [19]	83.8	2.4
SPOCO* [49]	84.4	1.7
BISSG* [32]	87.3	1.4
OGIS* [54]	87.5	1.1
OSFormer [37]	79.0	2.12
Mask2Former [6]	80.1	1.24
Ours	**88.7**	**0.7**

Table 2. Quantitative comparison with existing methods on the test set of CVPPP A1. $*$ denotes the updated results under the corrected calculation script of SBD on the CVPPP challenge website[1].

Results on CVPPP. We compare our method with existing methods on the test set of CVPPP A1. Since the ground truth labels of test data are not available, we report the results returned by the official challenge website. Due to the recent correction in the calculation script of SBD on the website, we reproduce the results of the latest methods with open source under the current calculation script and present them with * in Table 2. Note that the corrected calculation script would return a lower SBD value. As can be seen, compared with the two latest methods, PCTrans achieves the best SBD and |DiC| results. Specifically, there is an improvement of 1.2% SBD, which is the key metric of this dataset. Furthermore, we visualize the segmentation results on the validation set for comparison in the second example in Figure 3, which qualitatively demonstrates the superiority of our method. From the visualization, we can see that our PCTrans can precisely locate the objects and effectively segment the instance masks compared to other existing methods.

4.3.2 Comparison with Query-based Methoeds

In order to further illustrate the distinction between our PCTrans and existing query-based methods, we selected two representative works in natural images for comparison, including Mask2Former and OSFormer. As can be seen from Table 1 and 2, naively using queries to model biological instance feature results in severe performance drop due to high visual similarity. We also visualize the final segmentation results and the instance mask predicted by a single query in Figure 3 and 4, respectively. As shown in Figure 3, Mask2Former and OSFormer both suffer from more merge errors. The reason behind it can be explained from

[1] https://codalab.lisn.upsaclay.fr/competitions/8970

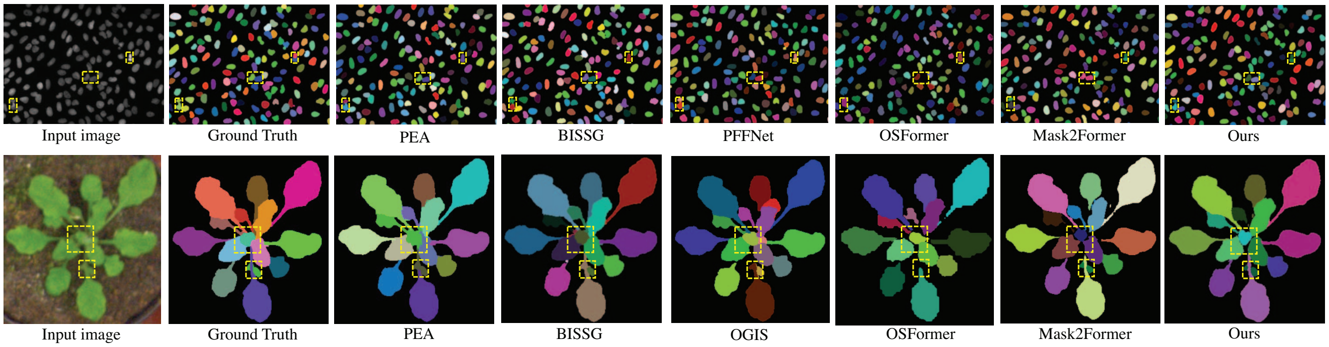

Figure 3. Visualization of segmentation results on the test set of BBBC039V1 (top) and the validation set of CVPPP (bottom). Different colors indicate different instances in the images. The yellow dashed boxes are drawn for clear comparison.

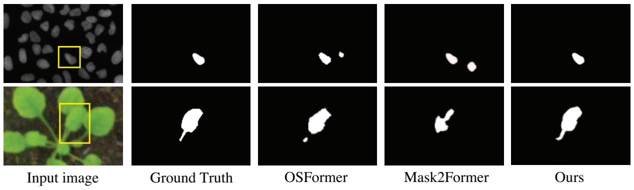

Figure 4. Quantitative comparison of mask candidates by queries in Msak2Former, OSFormer and Our PCTrans. The instance in yellow boxes are the target instances.

mechanism	AJI ↑	Dice ↑	F1 ↑	**PQ** ↑
Type 1	0.8588	0.9561	0.9474	0.8607
Type 2	0.8607	0.9550	0.9506	0.8640
Type 3	**0.8919**	**0.9647**	**0.9622**	**0.8833**

Table 3. Ablation results of different cross-attention mechanism.

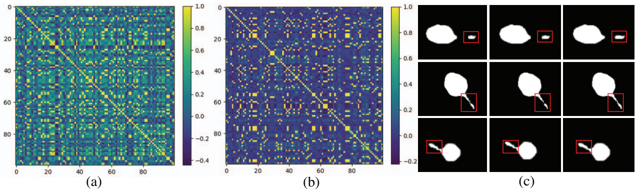

Figure 5. Visualization for the effect of the contrastive losses. (a) queries similarity w/o the contrastive losses, (b) queries similarity w/ the contrastive losses, (c) predicted masks by similar queries w/ the contrastive losses. The differences in predicted masks by similar queries are only litter in the red boxes.

Figure 4. The queries in Mask2Former and OSFormer are easier to predict worse masks and multiple instances once a time. Our PCTrans alleviates this problem to a large extent, which results in better segmentation performance.

4.4. Ablation Study and Analysis

To evaluate the effectiveness of key components adopted in our proposed PCTrans, we perform a series of comparisons with PCTrans variants on BBBC039V1. Table 4 shows the quantitative results with different components.

Effectiveness of position-guided cross-attention mechanism. The goal is to clarify how important the position-guided cross-attention mechanism is to our proposed PCTrans. We present the quantitative results of adopting the standard cross-attention (Type 1), the position-guided cross-attention with untransformed position embedding (Type 2) and the position-guided cross-attention with transformed position embedding (Type 3) in Table 3. As shown in Table 3, the position information cannot be utilized in the calculation of cross-attention without transforming position embedding conditioned on queries. Only by using queries to align the position embedding into a distinct embedding space, the position similarity can be considered in the cross-attention mechanism.

Effectiveness of position-aware dynamic mask head. The position information is also utilized in the dynamic mask head. The quantitative results with dynamic filters are shown in the first row and second row of Table 4. The results show that simply adopting dynamic filters without position information can provide stronger nonlinear properties to convert pixel embedding to instance masks. Besides, when concatenating the relative coordinate map into pixel embedding, the dynamic filters can be aware of the position information and better locate the distinct instance region. The ablation studies about position information indicate that position information is important when adapting queries to model instances features.

Effectiveness of two contrastive losses. Our proposed PCTrans aims to enhance the query representation by designing two contrastive losses. To verify the effectiveness, we compare the segmentation performance of our method with and without the QEC loss and the MCC loss. The results in Table 4 show that the contrastive losses have an important contribution to improving segmentation performance, which helps to enhance the segmentation masks and separate overlapping objects. We also visualize the query similarity and the predicted masks by similar queries for understanding the effect of these two contrastive losses in Figure 5. The results clearly show that the contrastive losses successfully suppress the feature similarity for different in-

dynamic filters	$\mathbf{M}_j^{rc}$	$\mathcal{L}_s$	$\mathcal{L}_d$	$\mathcal{L}_{\text{QEC}}$	$\mathcal{L}_{\text{MCC}}$	AJI ↑	Dice ↑	F1 ↑	**PQ** ↑
						0.7671	0.9601	0.8879	0.8011
✓						0.7684	0.9598	0.9091	0.8141
✓	✓					0.8591	0.9461	0.9424	0.8597
✓	✓	✓				0.8646	0.9599	0.9418	0.8611
✓	✓	✓	✓			0.8714	0.9482	0.9446	0.8665
✓	✓	✓	✓	✓		0.8921	0.9605	0.9570	0.8790
✓	✓	✓	✓		✓	0.8941	**0.9634**	0.9582	0.8821
✓	✓	✓	✓	✓	✓	**0.8981**	0.9617	**0.9615**	**0.8845**

Table 4. Ablation study of key components adopted in our proposed PCTrans. ✓ indicates adding the corresponding component.

τ_q	τ_m	AJI ↑	Dice ↑	F1 ↑	**PQ** ↑
0.5	0.5	**0.8983**	0.9625	0.9619	0.8856
0.5	1.0	0.8934	0.9619	0.9584	0.8816
0.5	2.0	0.8934	0.9620	0.9590	0.8813
1.0	0.5	0.8955	0.9632	0.9612	0.8851
1.0	1.0	0.8981	0.9617	0.9615	0.8845
1.0	2.0	0.8982	0.9632	0.9605	0.8844
2.0	0.5	0.8962	**0.9639**	**0.9621**	**0.8858**
2.0	1.0	0.8921	0.9626	0.9595	0.8816
2.0	2.0	0.8975	0.9624	0.9580	0.8805

Table 5. Ablation results of different temperature hyper-parameters on the test set of BBBC039V1.

Num. layers L	AJI ↑	Dice ↑	F1 ↑	**PQ** ↑
0	0.8752	0.9582	0.9422	0.8664
1	0.8696	0.9614	0.9512	0.8713
3	0.8848	0.9616	0.9530	0.8761
4	0.8883	0.9589	0.9574	0.8791
5	0.8955	0.9624	0.9575	0.8820
6	**0.8981**	0.9617	0.9615	**0.8845**
7	0.8923	**0.9618**	**0.9632**	0.8843

Table 6. Ablation results on the number of self-attention layer for extracting multi-scale pixel embeddings.

Dimension	AJI ↑	Dice ↑	F1 ↑	**PQ** ↑
4	0.8885	0.9588	0.9589	0.8764
8	0.8949	0.9607	0.9598	0.8780
16	**0.8981**	**0.9617**	**0.9615**	**0.8845**

Table 7. Ablation results for dimension of $\mathbf{F}_{mask}$.

Num. queries	AJI ↑	Dice ↑	F1 ↑	**PQ** ↑
200	0.8573	0.9446	0.9431	0.8671
300	**0.8981**	0.9617	0.9615	**0.8845**
600	0.8919	**0.9647**	**0.9622**	0.8833

Table 8. Ablation results on the number of queries.

stance queries and make queries more discriminative.

Temperature hyper-parameter. Table 5 shows the ablation results about different temperature hyper-parameters τ_q and τ_m in $\mathcal{L}_{\text{QEC}}$ and $\mathcal{L}_{\text{MCC}}$. Specifically, we use 0.5, 1.0 and 2.0 to produce different combinations of temperature hyperparameters τ_q and τ_m. As can be seen, $\tau_q = 2.0$ and $\tau_m = 0.5$ achieve the best performance on the PQ metric, which is the most important metric.

Number of self-attention layer L. The quality of multi-scale pixel embeddings is a key factor influencing the performance of the total network. We attempt a series of different numbers of self-attention layers for extracting pixel embedding to optimize the performance of PCTrans. As shown in Table 6, setting $L = 6$ is enough to get a good performance.

Dimension of $\mathbf{F}_{mask}$. We further investigate the impact of the $\mathbf{F}_{mask}$. We change C_{mask}, which is the number of channels of $\mathbf{F}_{mask}$. As shown in Table 7, higher dimensional embeddings are beneficial for better representation of pixels. $C_{mask} = 16$ is optimal and thus we use $C_{mask} = 16$ in all other experiments by default.

Number of queries. We study the number of queries on BBBC039V1 in Table 8. We can see that when $N = 300$, our method achieves the best result. However, there is an obvious performance degradation when $N = 200$. This suggests that a few queries are insufficient to provide a good result for the datasets with many instances in most biological images. The performance does not fluctuate much, although N becomes very large, suggesting that the mask candidates predicted by queries are robust.

5. Conclusion

In this paper, we propose the first one-stage query-based transformer PCTrans for biological instance segmentation. To address challenges such as high texture similarity and crowded objects, we incorporate position information into the learning process of queries and mask prediction. To overcome limited annotations, we propose two contrastive losses, QEC loss and MCC loss, to enhance the discriminative power of the queries representations. Experimental results demonstrate that our proposed PCTrans achieves state-of-the-art performance on two commonly used biological instance segmentation datasets.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62021001.

References

[1] Nikita Araslanov, Constantin A Rothkopf, and Stefan Roth. Actor-critic instance segmentation. In *CVPR*, 2019. 6

[2] Thorsten Beier, Constantin Pape, Nasim Rahaman, Timo Prange, Stuart Berg, Davi D Bock, Albert Cardona, Graham W Knott, Stephen M Plaza, Louis K Scheffer, et al. Multicut brings automated neurite segmentation closer to human performance. *Nature methods*, 14(2):101–102, 2017. 1, 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 4

[4] Vincent Casser, Kai Kang, Hanspeter Pfister, and Daniel Haehn. Fast mitochondria detection for connectomics. In *MIDL*, 2020. 1

[5] Qi Chen, Mingxing Li, Jiacheng Li, Bo Hu, and Zhiwei Xiong. Mask rearranging data augmentation for 3d mitochondria segmentation. In *MICCAI*, 2022. 1

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 4, 6

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 2

[8] Rupsa Datta, Tiffany M Heaster, Joe T Sharick, Amani A Gillette, and Melissa C Skala. Fluorescence lifetime imaging microscopy: fundamentals and advances in instrumentation, analysis, and applications. *Journal of biomedical optics*, 25(7):071203, 2020. 1

[9] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 2, 5, 6

[10] Daan De Geus, Panagiotis Meletis, and Gijs Dubbelman. Panoptic segmentation with a joint semantic and instance segmentation network. *arXiv preprint arXiv:1809.02110*, 2018. 6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2

[12] Jan Funke, Fabian Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas C Turaga. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1669–1680, 2018. 1, 2

[13] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. Sotr: Segmenting objects with transformers. In *ICCV*, 2021. 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[17] Hanzhe Hu, Jinshi Cui, and Liwei Wang. Region-aware contrastive learning for semantic segmentation. In *ICCV*, 2021. 3

[18] Wei Huang, Chang Chen, Zhiwei Xiong, Yueyi Zhang, Xuejin Chen, Xiaoyan Sun, and Feng Wu. Semi-supervised neuron segmentation via reinforced consistency learning. *IEEE Transactions on Medical Imaging*, 41(11):3016–3028, 2022. 1

[19] Wei Huang, Shiyu Deng, Chang Chen, Xueyang Fu, and Zhiwei Xiong. Learning to model pixel-embedded affinity for homogeneous instance segmentation. In *AAAI*, 2022. 5, 6

[20] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *NeurIPS*, 2016. 3

[21] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019. 6

[22] Victor Kulikov and Victor Lempitsky. Instance segmentation of biological images using harmonic embeddings. In *CVPR*, 2020. 2

[23] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 2

[24] Manan Lalit, Pavel Tomancak, and Florian Jug. Embedseg: Embedding-based instance segmentation for biomedical microscopy data. *Medical Image Analysis*, 81:102523, 2022. 1

[25] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *CVPR*, 2019. 6

[26] Dongnan Liu, Donghao Zhang, Yang Song, Heng Huang, and Weidong Cai. Panoptic feature fusion net: A novel instance segmentation paradigm for biomedical and biological images. *IEEE Transactions on Image Processing*, 30:2045–2059, 2021. 2, 6

[27] Dongnan Liu, Donghao Zhang, Yang Song, Chaoyi Zhang, Fan Zhang, Lauren O'Donnell, and Weidong Cai. Nuclei segmentation via a deep panoptic model with semantic feature fusion. In *IJCAI*, 2019. 2, 6

[28] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In *CVPR*, 2019. 6

[29] Xiaoyu Liu, Bo Hu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Efficient biomedical instance segmentation via knowledge distillation. In *MICCAI*, 2022. 2

[30] Xiaoyu Liu, Bo Hu, Mingxing Li, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. A soma segmentation benchmark in full adult fly brain. In *CVPR*, 2023. 1

[31] Xiaoyu Liu, Wei Huang, Zhiwei Xiong, Shenglong Zhou, Yueyi Zhang, Xuejin Chen, Zheng-Jun Zha, and Feng Wu. Learning cross-representation affinity consistency for sparsely supervised biomedical instance segmentation. In *ICCV*, 2023. 2

[32] Xiaoyu Liu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Biological instance segmentation with a superpixel-guided graph. In *IJCAI*, 2022. 6

[33] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012. 1, 5

[34] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *ICCV*, 2021. 4

[35] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 4

[36] Jean-Michel Pape and Christian Klukas. 3-d histogram-based segmentation and leaf detection for rosette plants. In *ECCV*, 2014. 6

[37] Jialun Pei, Tianyang Cheng, Deng-Ping Fan, He Tang, Chuanbo Chen, and Luc Van Gool. Osformer: One-stage camouflaged instance segmentation with transformers. In *ECCV*, 2022. 2, 4, 6

[38] Hanno Scharr, Massimo Minervini, Andreas Fischbach, and Sotirios A Tsaftaris. Annotated image datasets of rosette plants. In *ECCV*, 2014. 1, 5

[39] Hanno Scharr, Massimo Minervini, Andrew P French, et al. Leaf segmentation in plant phenotyping: a collation study. *Machine Vision and Applications*, 27(4):585–606, 2016. 6

[40] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *CVPR*, 2016. 3

[41] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020. 3, 4

[42] Eric Upschulte, Stefan Harmeling, Katrin Amunts, and Timo Dickscheid. Contour proposal networks for biomedical instance segmentation. *Medical image analysis*, 77:102371, 2022. 1

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[44] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, page 53, 2019. 1

[45] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 2021. 2

[46] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *ICCV*, 2021. 3

[47] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. In *NeurIPS*, 2020. 3

[48] Steffen Wolf, Alberto Bailoni, Constantin Pape, Nasim Rahaman, Anna Kreshuk, Ullrich Köthe, and Fred A Hamprecht. The mutex watershed and its objective: Efficient, parameter-free graph partitioning. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3724–3738, 2020. 1, 2

[49] Adrian Wolny, Qin Yu, Constantin Pape, and Anna Kreshuk. Sparse object-level supervision for instance segmentation with pixel embeddings. In *CVPR*, 2022. 2, 6

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 2

[51] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, 2019. 6

[52] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. In *NeurIPS*, 2019. 3

[53] Jingru Yi, Hui Tang, Pengxiang Wu, Bo Liu, Daniel J Hoeppner, Dimitris N Metaxas, Lianyi Han, and Wei Fan. Object-guided instance segmentation for biological images. In *AAAI*, 2020. 2

[54] Jingru Yi, Pengxiang Wu, Hui Tang, Bo Liu, Qiaoying Huang, Hui Qu, Lianyi Han, Wei Fan, Daniel J Hoeppner, and Dimitris N Metaxas. Object-guided instance segmentation with auxiliary feature refinement for biological images. *IEEE Transactions on Medical Imaging*, 40(9):2403–2414, 2021. 2, 6

[55] Donghao Zhang, Yang Song, Dongnan Liu, Haozhe Jia, Siqi Liu, Yong Xia, Heng Huang, and Weidong Cai. Panoptic segmentation with an end-to-end cell r-cnn for pathology image analysis. In *MICCAI*, 2018. 6

[56] Qilong Zhangli, Jingru Yi, Di Liu, Xiaoxiao He, Zhaoyang Xia, Qi Chang, Ligong Han, Yunhe Gao, Song Wen, Haiming Tang, et al. Region proposal rectification towards robust instance segmentation of biological images. In *MICCAI*, 2022. 2

[57] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 5