2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

Attention-Based Transformers for Instance Segmentation of Cells in Microstructures

Tim Prangemeier, Christoph Reich, Heinz Koeppl‡
Centre for Synthetic Biology,
Department of Electrical Engineering and Information Technology, Department of Biology,
Technische Universität Darmstadt
‡*heinz.koeppl@bcs.tu-darmstadt.de*

***Abstract*—Detecting and segmenting object instances is a common task in biomedical applications. Examples range from detecting lesions on functional magnetic resonance images, to the detection of tumours in histopathological images and extracting quantitative single-cell information from microscopy imagery, where cell segmentation is a major bottleneck. Attention-based transformers are state-of-the-art in a range of deep learning fields. They have recently been proposed for segmentation tasks where they are beginning to outperform other methods. We present a novel attention-based *cell detection transformer* (Cell-DETR) for direct end-to-end instance segmentation. While the segmentation performance is on par with a state-of-the-art instance segmentation method, Cell-DETR is simpler and faster. We showcase the method's contribution in a the typical use case of segmenting yeast in microstructured environments, commonly employed in systems or synthetic biology. For the specific use case, the proposed method surpasses the state-of-the-art tools for semantic segmentation and additionally predicts the individual object instances. The fast and accurate instance segmentation performance increases the experimental information yield for *a posteriori* data processing and makes online monitoring of experiments and closed-loop optimal experimental design feasible. Code and data samples are available at https://git.rwth-aachen.de/bcs/projects/cell-detr.git.**

***Index Terms*—attention, instance segmentation, transformers, single-cell analysis, synthetic biology, microfluidics, deep learning**

I. Introduction

Instance segmentation is a common task in biomedical applications. It is comprised of both detecting individual object instances and segmenting them [1], [2]. Prevalent examples in healthcare and life sciences include the detection of individual tumour or cell entities and the segmentation of their shape. Recent advances in automated single-cell image processing, such as instance segmentation, have contributed to early tumour detection, personalised medicine, biological signal transduction and insight into the mechanisms behind cell population heterogeneity, amongst others [3]–[6]. An example of instance segmentation is shown in Fig. 1, with four separate cell and two trap microstructures detected and segmented individually.

Object detection and panoptic segmentation are closely related to instance segmentation [2], [7]. Carion *et al.* recently proposed a novel attention-based *detection transformer* DETR for panoptic segmentation [8]. DETR achieves state-of-the-art panoptic segmentation performance, while exhibiting a comparatively simple architecture that is easier to implement and is computationally more efficient than previous approaches [8]. Its simplicity promises to be beneficial for its adoption in real-world applications.

Time-lapse fluorescence microscopy (TLFM) is a powerful technique for studying cellular processes in living cells [4], [9]–[11]. The vast amount of quantitative data TLFM yields, promises to constitute the backbone of the rational design of *de novo* biomolecular functionality [10], [11]. Ideally in synthetic biology, well characterised parts are combined *in silico* in a quantitatively predictive bottom up approach [11]–[13], for example, to detect and kill cancer cells [14], [15].

Quantitative TLFM with high-throughput microfluidics is an essential technique for concurrently studying the heterogeneity and dynamics of synthetic circuitry on the single cell level [4], [9], [11]. A typical TLFM experiment yields thousands of specimen images (Fig. 1) requiring automated segmentation, examples include [5], [16], [17]. Segmenting each individual cell enables its pertinent information to be extracted quantitatively. For example, the abundance of a fluorescent reporter can be measured, giving insight into the cell's inner workings.

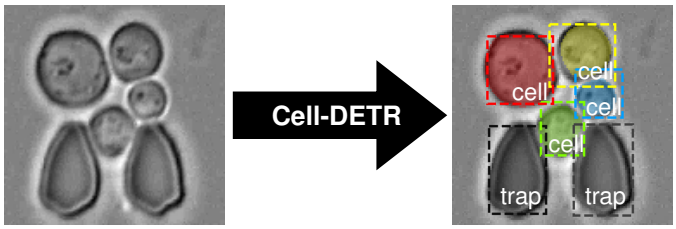

Fig. 1. Schematic of Cell-DETR direct instance segmentation discerning individual cell (colour) and trap microstructure (grey) object instances.

Instance segmentation is a major bottleneck in quantifying single-cell microscopy data and manual analysis is prohibitively labour intensive [9], [11], [16], [18], [19]. The vast majority of single-cell segmentation methods are designed for *a posteriori* data processing and often require post-processing for instance detection or manual input [9]. This is not only a drawback on the amount of experiments that can be performed, but also limits the type of experiments [18], [20]. For example, harnessing the potential of advanced closed-loop optimal experimental design techniques [12], [21], [22] requires online monitoring with fast instance segmentation

978-1-7281-6215-7/20/$31.00 ©2020 IEEE
Urheberrechtlich geschützt / In Copyright

© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

capabilities. Attention-based methods, such as the recently proposed detection transformer DETR [8], are increasingly outperforming other methods [8], [23]. For the yeast-trap configuration (Fig. 1) direct instance segmentation has yet to be employed and attention-based transformers have yet to be applied for segmentation in the biomedical fields in general.

In this study, we present Cell-DETR, a novel attention-based detection transformer for instance segmentation of biomedical samples based on DETR [8]. We address the automated cell instance segmentation bottleneck for yeast cells in microstructured environments (Fig. 1) and showcase Cell-DETR on this application. Section II introduces the previous segmentation approaches and the microstructured environment. Our experimental setup for fluorescence microscopy, the tested architectures and our approach to training and evaluation are presented in Section III. We analyse the proposed method's performance in Section IV and compare it to the application specific state-of-the-art, as well as to a general instance segmentation baseline. After interpreting the results and highlighting the method's future potential in Section V, we summarise and conclude the study in Section VI. Our model surpasses the previous application baseline and is on par with a general state-of-the-art instance segmentation method. The relatively short inference runtimes enable higher throughput *a posteriori* data processing and make online monitoring of experiments with approximately 1000 cell traps feasible.

II. Background

An extensive body of research into the automated processing of microscopy imagery dates back to the middle of the 20-th century. Recent studies demonstrate the utility of deep learning segmentation approaches, for example [6], [9], [19], [24], [25]. Comprehensive reviews of the many methods to segment yeast on microscopy imagery are available elsewhere [3], [20]. Here we focus on dedicated tools for segmenting cells in trapped microstructures. U-Net convolutional neural networks (CNNs) with an encoder-decoder architecture bridged by skip connections have been shown to perform semantic segmentation well for *E. coli* mother machines [9], [19] and yeast in microstructured environments [6]. In the case of trapped yeast, the previous state-of-the-art tool DISCO [16] was based on conventional methods (template matching, support vector machine, active contours), until recently being superseded by U-Nets [6]. The current baseline for semantic segmentation of yeast in microstructured environments, as measured by the cell class intersection-over-union, is 0.82 [6]. Additional post-processing, of the segmentation maps is required to attain each individual cell instance [6].

For instance segmentation in general, recent state-of-the-art methods are available, for example Mask R-CNN [1]. It is a proposal-based instance segmentation model, which combines a CNN backbone, region proposals with non-maximum-suppression, region-of-interest (ROI) pooling, and multiple prediction heads [1]. Attention-based methods are increasingly outperforming convolutional methods and are currently state-of-the-art in natural language processing [23]. Beyond

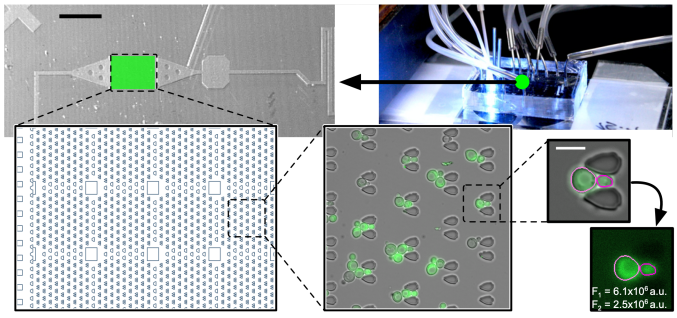

Fig. 2. Single-cell fluorescence measurement setup. Microfluidic chip on the microscope table (top right), microscope imagery and design of the yeast trap microstructures. The trap chamber (green rectangle) contains an array of approximately 1000 traps. Single specimen images show a pair of microstructures and fluorescent cells, violet contours indicate segmentation of two separate cell instances with corresponding fluorescence measurement F_1 and F_2; black scale bar 1 mm, white scale bar 10 µm.

natural language processing, attention-based approaches, such as axial-attention modules [26], have demonstrated promising results in computer vision applications [8]. Recently, the first transformer-based method (DETR [8]) for object detection and panoptic segmentation was reported. DETR achieves state-of-the-art results on par with Faster R-CNN and constitutes a promising approach for further improvements in automated object detection and segmentation performance.

The microfluidic trap microstructures we consider here are designed for long-term culture of yeast cells (*Saccharomyces cerevisiae*) within the focal plane of a microscope [17]. The on-chip environment is tightly controlled and conducive to yeast growth. Examples of its routine employ include Fig. 2 and [4]–[6], [11], [16]. A constant flow of yeast growth media hydrodynamically traps the cells in the microstructures and allows the introduction of chemical perturbations. An automated microscope records an entire trap chamber of up to 1000 traps by imaging both the brightfield and fluorescent channels at approximately 20 neighbouring positions. Typical experiments each produce hundreds of GB of image data. Time-lapse recordings allow individual cells to be tracked through time. Robust instance segmentation facilitates tracking [9], [20], which itself can be a limiting factor with regard to the data yield of an experiment [4], [9], [19], [20].

III. Methodology

A. Live-cell microscopy dataset and annotations

The individual specimen images each contain a single microfluidic trap and some yeast cells, as depicted in Fig. 1. These are extracted from larger microscope recordings, whereby each exposure contains up to 50 traps (Fig. 2 middle). Ideally, a single mother cell persists in each trap, with subsequent daughter cells being removed by the constant flow of yeast growth media. In practice, multiple cells accumulate around some traps, while other traps remain empty (Fig. 3).

We distinguish between three classes on the specimen image annotations, as depicted in Fig. 3. The yeast cells in violet are the most important class for biological applications. To

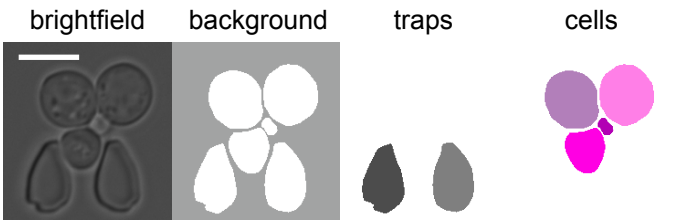

Fig. 3. Example of class and instance annotations for a specimen image; brightfield image (left), background label in light grey, instances of the trap class in shades of dark grey and instances of the cell class in shades of violet (left to right respectively); scale bar 10 μm.

counteract traps being segmented as cells, we employ a distinct class for them (dark grey). The background (light grey) is annotated for semantic segmentation training, for example of U-Nets. For instance segmentation training we introduce a *no-object* class $\varnothing$ in place of the background class.

Each instance of cells or trap structures are annotated individually with a bounding box, class specification and separate pixel-wise segmentation map. Here we omit the bounding boxes to enable an unobscured view of the contours. Instead, the distinct cell instances and their individual segmentation maps are indicated by different shades of violet in Fig. 3.

The annotated set of 419 specimen images from various experiments was randomly assigned for network training, validation and testing (76 %, 12 % and 12 % respectively). Examples are shown in Fig. 4. Images include a balance of the common yeast-trap configurations: 1) empty traps, 2) single cells (with daughter) and 3) multiple cells. Slight variations in trap fabrication, debris, contamination, focal shift, illumination levels and yeast morphology were included. Further scenarios or strong variations, such as trap design geometries, model organisms and significant focal shift, were omitted.

B. The Cell-DETR instance segmentation architecture

The proposed Cell-DETR models A and B are based on the DETR panoptic segmentation architecture [8]. We adapted the architecture for non-overlapping instance segmentation and reduced it in size for faster inference. The main differences between DETR and our variants Cell-DETR A and B are summarised in Table I. The Cell-DETR variants have approximately one order of magnitude less parameters than the original ($\sim 40 \times 10^6$ reduced to $\sim 5 \times 10^6$ parameters). The main building blocks of the Cell-DETR model are detailed in Fig. 5. They are the backbone CNN encoder, the transformer encoder-decoder, the bounding box and class prediction heads, and the segmentation head.

The CNN encoder (left in Fig. 5) extracts image features of the brightfield specimen image input. It is based on four ResNet-like [27] blocks with 64, 128, 256 and 256 convolutional filters. After each block a 2×2 average pooling layer is utilised to downsample the intermediate feature maps. The Cell-DETR variants employ different activations and convolutions, as detailed in Table I.

The transformer encoder determines the attention between image features. The transformer decoder predicts the attention

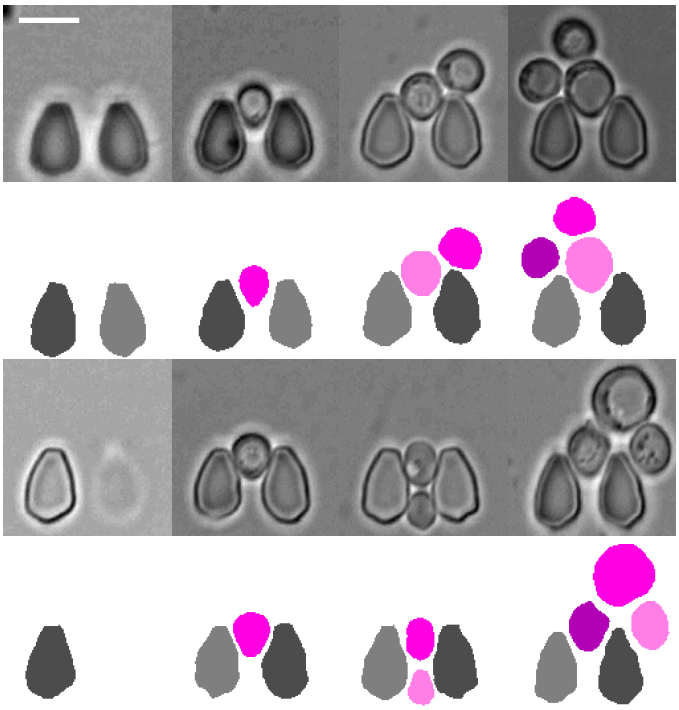

Fig. 4. Characteristic selection of specimen images and corresponding annotations, including empty or single trap structures, trapped single cells (with single daughter adjacent) and multiple trapped cells; trap instances in shades of dark grey, cell instances in shades of violet and transparent background; scale bar 10 μm.

regions for each of the $N = 20$ object queries. They are both based on the DETR architecture [8]. We reduced the number of transformer encoder blocks to three and decoder blocks to two, each with 512 hidden features in the feed-forward neural-network (FFNN). The 128 backbone features are flattened before being fed into the transformer. In contrast to the original DETR, we employed learned positional encodings. While Cell-DETR A employs leaky ReLU [28] activations, Padé activation units [29] are utilised for Cell-DETR B.

TABLE I
OVERVIEW OF DIFFERENCES BETWEEN DETR, CELL-DETR A AND B.

Model	Activation functions	Convolutions	Feature fusion	Param. $\times 10^6$
DETR [8]	ReLU	standard spatial	addition	$\gtrapprox 40$
C-DETR A	leaky ReLU [28]	standard spatial	addition	4.3
C-DETR B	Padé [29]	deformable (v2) [30]	pix.-adapt. conv. [31]	5.0

The prediction heads for the bounding box and classification are each a FFNN. They map the transformer encoder-decoder output to the bounding box and classification prediction. These FFNN process each query in parallel and share parameters over all queries. In addition to the cell and trap classes, the classification head can also predict the no-object class $\varnothing$.

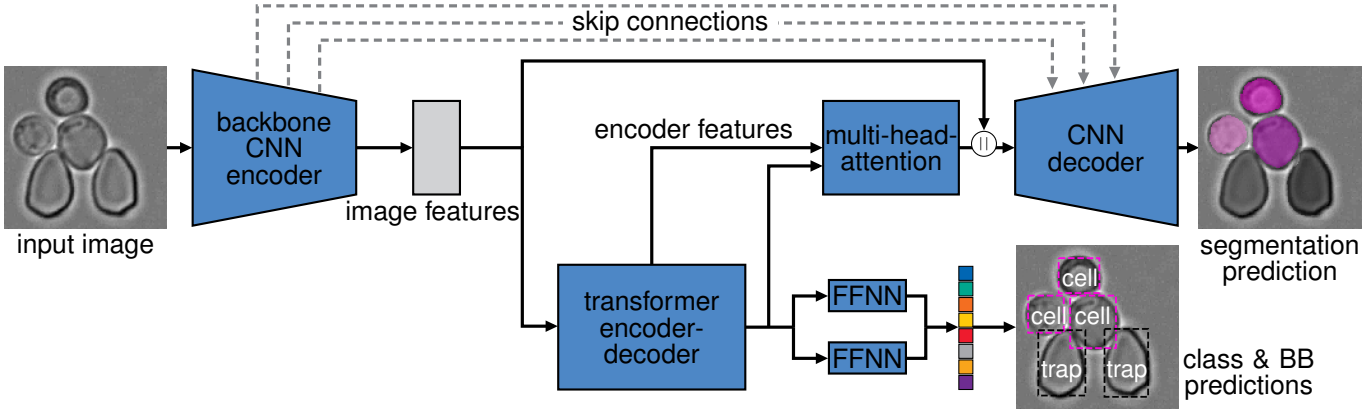

Fig. 5. Architecture of the end-to-end instance segmentation network, with brightfield specimen image input and an instance segmentation prediction as output. The backbone CNN encoder extracts image features that then feed into both the transformer encoder-decoder for class and bounding box prediction, as well as to the CNN decoder for segmentation. The transformer encoded features, as well as the transformer decoded features, are fed into a multi-head-attention module and together with the image features from the CNN backbone feed into the CNN decoder for segmentation. Skip connections additionally bridge between the backbone CNN encoder and the CNN decoder. Input and output resolution is 128 × 128 pixels.

The segmentation head is composed of a multi-head attention mechanism and a CNN decoder to predict the segmentation maps for each object instance. We employ the original DETR [8] two-dimensional multi-head attention mechanism between the transformer encoder and decoder features. The resulting attention maps are concatenated channel-wise onto the image features and fed into the CNN decoder. The three ResNet-like decoder blocks decrease the feature channel size while increasing the spatial dimensions. Long skip connections bridge between the CNN encoder and CNN decoder blocks' respective outputs. The features are fused by element-wise addition in Cell-DETR A and by pixel-adaptive convolutions in Cell-DETR B. A fourth convolutional block incorporates the queries in the feature dimension and returns the original input's spatial dimension for each query. Non-overlapping segmentation is ensured by a softmax over all queries.

C. Training Cell-DETR

We employ a combined loss function and a direct set prediction to train our Cell-DETR networks end-to-end. The set prediction $\hat{y} = \{\hat{y}_i = \{\hat{p}_i, \hat{b}_i, \hat{s}_i\}\}_1^{N=20}$ is comprised of the respective predictions for class probability $\hat{p}_i \in \mathbb{R}^K$ (here $K = 3$ classes, no-object, trap, cell), bounding box $\hat{b}_i \in \mathbb{R}^4$ and segmentation $\hat{s}_i \in \mathbb{R}^{128\times128}$ for each of the N queries. We assigned each instance set label $y_{\sigma(i)}$ to the corresponding query set prediction $\hat{y}_i$ with the Hungarian algorithm [8], [32]. The indices σ_i denote the best matching permutation of labels.

The combined loss $\mathcal{L}$ is comprised of a classification loss $\mathcal{L}_p$, a bounding box loss $\mathcal{L}_b$, and a segmentation loss $\mathcal{L}_s$

$$\mathcal{L} = \sum_{i=1}^{N} \left(\mathcal{L}_p + \mathbb{1}_{\{p_i \neq \varnothing\}} \mathcal{L}_b + \mathbb{1}_{\{p_i \neq \varnothing\}} \mathcal{L}_s \right),$$

with $N = 20$ object instance queries in this case. We employ class-wise weighted cross entropy for the classification loss

$$\mathcal{L}_p \left(p_{\sigma(i)}, \hat{p}_i \right) = -\sum_{k=1}^{K} \beta_k \, p_{\sigma(i),k} \log(\hat{p}_{i,k}),$$

with weights $\beta = [0.5, 0.5, 1.5]$ for the $K = 3$ classes, no-object, trap and cell classes respectively. The bounding box loss is itself composed of two weighted loss terms. These are a *generalised intersection-over-union* $\mathcal{L}_\mathcal{J}$ [33], and a L1 loss, with respective weights $\lambda_\mathcal{J} = 0.4$ and $\lambda_{L1} = 0.6$

$$\mathcal{L}_b \left(b_{\sigma(i)}, \hat{b}_i \right) = \lambda_\mathcal{J} \, \mathcal{L}_\mathcal{J} \left(b_{\sigma(i)}, \hat{b}_i \right) + \lambda_{L1} \left|\left| b_{\sigma(i)} - \hat{b}_i \right|\right|_1 .$$

The segmentation loss $\mathcal{L}_s$ is a weighted sum of the focal loss $\mathcal{L}_\mathcal{F}$ [34] and Sørensen-Dice loss $\mathcal{L}_D$ [6], [8]

$$\mathcal{L}_s \left(s_{\sigma(i)}, \hat{s}_i \right) = \lambda_\mathcal{F} \, \mathcal{L}_\mathcal{F} \left(s_{\sigma(i)}, \hat{s}_i; \gamma \right) + \lambda_\mathcal{D} \, \mathcal{L}_\mathcal{D} \left(s_{\sigma(i)}, \hat{s}_i; \epsilon \right).$$

The respective weights are $\lambda_\mathcal{F} = 0.05$ and $\lambda_\mathcal{D} = 1$, with focusing parameter $\gamma = 2$ and $\epsilon = 1$ for numerical stability.

D. Evaluation and implementation

We employ a number of metrics to quantitatively analyse the performance of the trained networks with regard to classification, bounding box and segmentation performance. Given the ground truth $\mathbf{Y}$ and the prediction $\hat{\mathbf{Y}}$ (in the corresponding instance-matched permutation), we evaluate the segmentation performance with variants of the Jaccard index $\mathcal{J}$ and the Sørensen-Dice $\mathcal{D}$ coefficient [6], [8], omitting the background

$$\mathcal{D}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{2|\mathbf{Y} \cap \hat{\mathbf{Y}}|}{|\mathbf{Y}| + |\hat{\mathbf{Y}}|}; \quad \mathcal{J}_k(\mathbf{Y}_k, \hat{\mathbf{Y}}_k) = \frac{|\mathbf{Y}_k \cap \hat{\mathbf{Y}}_k|}{|\mathbf{Y}_k \cup \hat{\mathbf{Y}}_k|}, \quad (1)$$

with $\mathcal{J}_k$ intuitively the *intersection-over-union* for each class k. With respect to the metrological application in image cytometry, the cell class is of most importance, therefore, we consider the Jaccard index for the cell class alone ($\mathcal{J}_c$). Similarly, in the case of instance segmentation, we compute $\mathcal{J}_i$ for each instance i and average over all I object instances to compute the mean instance Jaccard index $\bar{\mathcal{J}}_I = \frac{1}{I} \sum_{i=1}^{I} \mathcal{J}_i$.

We utilise the accuracy as the proportion of correct predictions for classification. The bounding boxes are evaluated with the Jaccard index $\bar{\mathcal{J}}_b$. It is defined analogously to the object instance Jaccard index (compare Eqn. 1), yet computed implicitly with the bounding box coordinates.

We compare the proposed method with our own implementations of both the state-of-the-art for the trapped yeast application (U-Net [6]), as well as more generally with a state-of-the-art instance segmentation meta algorithm (Mask R-CNN [1]). The multiclass U-Net for semantic segmentation was implemented in PyTorch, with the architecture, pre- and post-processing described in [6]. We implemented a Mask R-CNN [1] with Torchvision (PyTorch) and a ResNet-18 [27] backbone, which was pre-trained for image classification.

We implemented the proposed Cell-DETR A and B architectures with PyTorch. We used the MMDetection toolkit [35] for deformable convolutions and the PyTorch/Cuda implementation for the Padé activation units [29]. We trained the models using AdamW [36] for optimisation with a weight decay of 10^{-6}. The initial learning rate was 10^{-5} for the backbone and 10^{-4} for the rest of the model. The learning rates were decreased by an order of magnitude after 50 and again 100 epochs of the total 200 epochs. The additional first and second-order momentum moving average factors were 0.9 and 0.999 respectively. We selected the best performing model based on the cell class Jaccard index $\mathcal{J}_c$, typically after 80 to 140 epochs with mini batch size 8. The training data was randomly augmented by elastic deformation [6], [24], horizontal flipping or by the addition of noise with a probability of 0.6. Inference runtimes for one forward pass were averaged over 1000 runs on a Nvidia RTX 2080 Ti for all three methods (U-Net, Mask R-CNN and Cell-DETR).

E. Data acquisition setup

Yeast cells were cultured in a tightly controlled microfluidic environment. A temperature of $30\,^{\circ}\mathrm{C}$ and the flow of yeast growth media enables yeast to grow for prolonged periods and over multiple cell-cycles. The microfluidic chips confined the cells to the focal plane of the microscope. Continuous media flow hydrodynamically traps the living cells in the microstructures. The Polydimethylsiloxane (PDMS) microstructures constrain the cells in XY, while axial constraints in Z are provided by the cover slip and PDMS ceiling. The space between cover slip and the PDMS ceiling is on the order of a cell diameter to facilitate continuously uniform focus of the cells.

We recored time-lapse brightfield (transmitted light) and fluorescent channel imagery of the budding yeast cells every $10\,\mathrm{min}$ with a computer controlled microscope (Nikon Eclipe Ti with XYZ stage; μManager; 60x objective). A CoolLED pE-100 and a Lumencor SpectraX light engine illuminated the respective channels, which were captured with a ORCA Flash 4.0 (Hamamatsu) camera. Multiple lateral and axial positions were recorded sequentially at each timestep (Fig. 2).

IV. Results

A. Cell-DETR variant results

A sample of segmentation results for the two Cell-DETR variants is shown in Fig. 6. The cell and trap instances are all detected and classified correctly with slight variations in segmentation contours. Separate instances of cells and traps are indicated by the shades of violet and grey respectively. Variant B demonstrates slightly better segmentation performance. A qualitative example of this is shown in in Fig. 6, where Cell-DETR A in contrast to B excludes a small section of one cell.

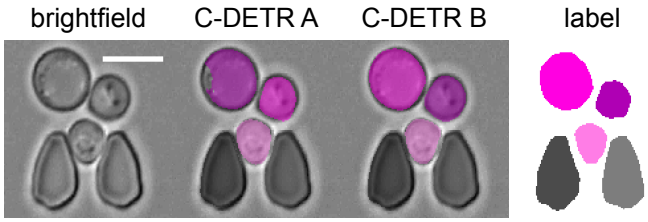

Fig. 6. Qualitative comparison of Cell-DETR A and B segmentation examples for a selected test image (left) and label (right); trap instances in shades of dark grey, cell instances in shades in violet; scale bar $10\,\mu\mathrm{m}$.

The quantitative comparison between the segmentation performance of the Cell-DETR variants is summarised in Table II. We modified model B for better performance on our application, as described in Section III-B. The mean Jaccard index over all object instances increased from $\bar{\mathcal{J}}_I = 0.84$ for model A to $\bar{\mathcal{J}}_I = 0.85$ for model B, while the cell class Jaccard index increased by a similar margin from $\mathcal{J}_c = 0.83$ to $\mathcal{J}_c = 0.84$. Taking the background into account, a segmentation accuracy of 0.96 is achieved. Both Cell-DETR surpass the segmentation performance ($\mathcal{J}_c$) of the previous state-of-the-art methods for the trapped yeast application [6], [16], in addition to directly attaining the instances.

TABLE II
Segmentation performance of Cell-DETR A and B.

Model	Sørensen Dice $\mathcal{D}$	Mean instance $\bar{\mathcal{J}}_I$	Cell class $\mathcal{J}_c$	Seg. accuracy
C-DETR A	0.92	0.84	0.83	0.96
C-DETR B	0.92	0.85	0.84	0.96

The bounding box and classification performance is summarised in Table III. Again, both models perform similarly well. They correctly classify the object instances (classification accuracy of 1.0) and detect the correct number of instances for each class. They also perform similarly well at localising the instances, achieve a bounding box *intersection-over-union* of $\mathcal{J}_b = 0.81$, for the standard formulation as well as the generalised form employed for training.

TABLE III
Bounding box and classification performance metrics for Cell-DETR A and B.

Model	Bounding box Jaccard $\mathcal{J}_b$	Classification accuracy
C-DETR A	0.81	1.0
C-DETR B	0.81	1.0

The slight increase in segmentation performance that model B yields is a trade off with increased computational cost. The number of parameters is increased from approximately 4×10^6, to over 5×10^6 (Table I). This leads to an increase in runtime

from 9.0 ms for model A to 21.2 ms for model B. These times are orders of magnitude faster than the previous state-of-the-art method DISCO [16] and on the same order of magnitude as the currently fastest reported network for this application [6]. Runtimes on this order of magnitude suffice for in-the-loop experimental techniques.

We select model B for further analysis, based on the improved performance and sufficiently fast runtimes. A selection of segmentation predictions for the three most typical scenarios in the test dataset is given in Fig. 7. The detection of cell and trap instances, without any overlap between instances, is successful for single cells (middle row), multiple cells (bottom row), and empty traps are correctly identified. The introduction of multiple classes (traps, cells), as well as individual object instances facilitated individually segmenting each cell entity and discerning these from both the traps and other cells.

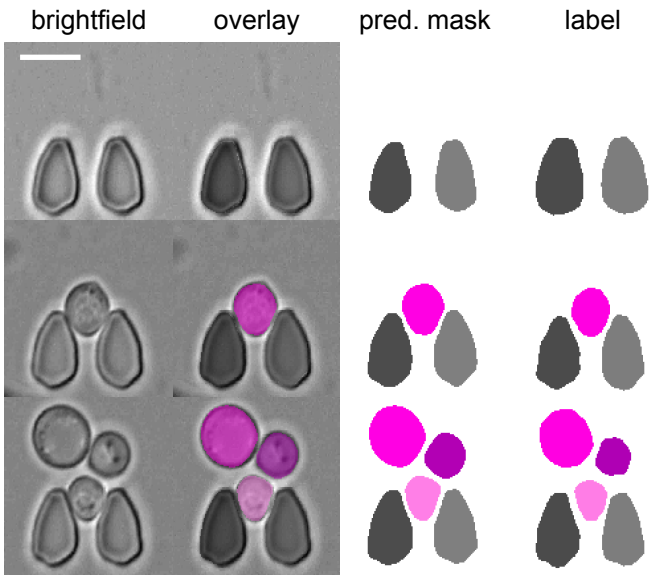

Fig. 7. Example of different scenarios from the test dataset segmented with Cell-DETR B: an empty trap (top row), a single trapped cell (middle row) and multiple cells; columns are brightfield, an overlay of the prediction, the prediction mask and the ground truth label (left to right respectively). Colours indicate traps in shades of grey ■ and cell instances in shades of violet ■; scale bar 10 μm.

The intended application of our method is to deliver segmentation masks for each cell instance for subsequent single-cell fluorescence measurements. We trialled this application on unlabelled and unseen data as depicted in Fig. 8. The cell instances are detected based on the brightfield image (left) and the resulting object segmentation predictions are used as masks to measure the individual cell fluorescence on the fluorescent channel (right). An overlay of the brightfield, fluorescent images with the segmentation contours is depicted in the middle, along with the green fluorescent protein (GFP) channel. The individual cell area (A_1 and A_2) is measured as the number of pixels in the instance segmentation mask and indicated on the GFP channel. The cell instance fluorescence (F_1 and F_2) is summed over the mask area and indicated on the right for each individual cell in arbitrary fluorescence units.

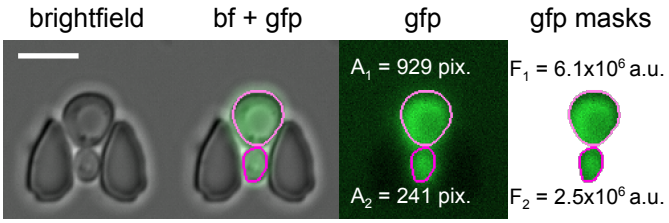

Fig. 8. Example of individual cell fluorescence measurement application with a segmentation mask contour for each individual cell (violet contours violet ■) based on the brightfield image (left); scale bar 10 μm.

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

We compare our proposed method with the state-of-the-art for the trapped yeast application (DISCO [16], U-Net [6]), as well as with a general state-of-the-art method for instance segmentation (Mask R-CNN [1]). We implemented both the U-Net and Mask R-CNN methods in this study (Section III-D). A characteristic qualitative example of the results is given in Fig. 9, with the ground truth on the left, followed by Cell-DETR B, Mask R-CNN and U-Net segmentations results. All three methods segment two trap microstructures and all four cells in separate classes, without any overlap or touching cells. Cell-DETR B and Mask R-CNN additionally segment each cell or trap object as an individual instance. The contours are slightly smaller for the U-Net, which is deemed a result of the emphasis on avoiding touching cells and the associated difficulty of discerning these in subsequent post-processing.

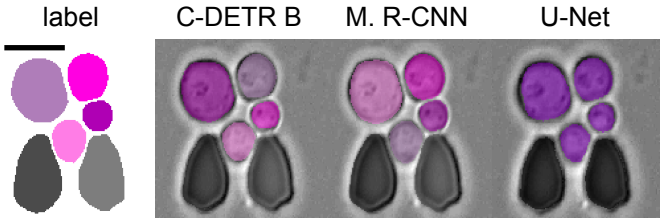

Fig. 9. Example segmentation for our implementations of Cell-DETR B, Mask R-CNN and U-Net. Trap instances in shades of grey ■ and cell instances in shades of violet ■ (no instance detection for U-Net); scale bar 10 μm.

Accurate segmentation of the cells is particularly important for the measurement of cell morphology or fluorescence. We compare the cell class Jaccard index $\mathcal{J}_c$ of our proposed methods Cell-DETR A and B with the application state-of-the-art methods DISCO, U-Net and Mask R-CNN. The comparison is summarised in Table IV. U-Net recently superseded DISCO [16] ($\mathcal{J}_c \sim 0.7$) as the state-of-the-art trapped yeast segmentation method, achieving $\mathcal{J}_c = 0.82$. Our Cell-DETR variants both further improve on this result, with model B achieving the same $\mathcal{J}_c = 0.84$ on par with our Mask R-CNN implementation. Cell-DETR and Mask R-CNN additionally provide each cell object instance.

We measured the average runtime of a forward pass of each method on a single specimen image (Table IV). For DISCO

TABLE IV
COMPARISON OF CELL-DETR PERFORMANCE WITH THE STATE-OF-THE-ART METHODS FOR THE TRAPPED YEAST APPLICATION (DISCO, U-NET) AND INSTANCE SEGMENTATION (MASK R-CNN).

Model	Cell Class $\mathcal{J}_c$	Inference runtime[1]	Instances
DISCO [16][2]	~ 0.70	$\sim 1300\,\mathrm{ms}$	×
U-Net	0.82	$1.8\,\mathrm{ms}$	×
Mask R-CNN	0.84	$29.8\,\mathrm{ms}$	✓
Cell-DETR A	0.83	$9.0\,\mathrm{ms}$	✓
Cell-DETR B	0.84	$21.2\,\mathrm{ms}$	✓

[1] Runtimes for U-Net, Mask R-CNN, and Cell-DETR averaged over 1000 runs ($\sim$ 300 different images) on a Nvidia RTX 2080 Ti; measurement uncertainty is below $\pm 5\%$.
[2] Reported literature values [16].

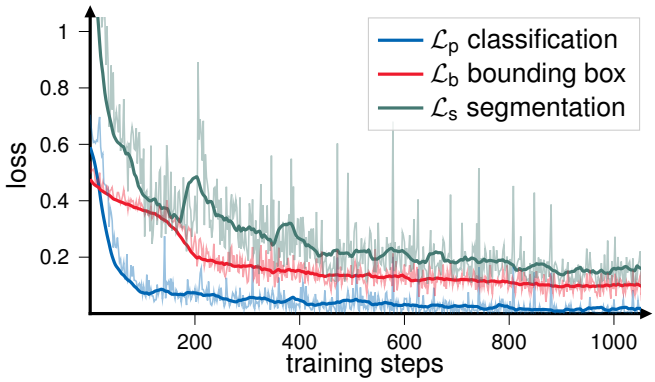

Fig. 10. Classification, bounding boxes and segmentation loss curves for Cell-DETR B; thick lines are running averages (window size 30).

[16] we consider the reported values that include some pre- and post-processing steps to detect cells individually. The deep methods are significantly faster than DISCO, making online monitoring of live experiments feasible. The U-Net is the fastest, taking $1.8\,\mathrm{ms}$ for a forward pass, in contrast to $29.8\,\mathrm{ms}$ for the Mask R-CNN [1]. However, the U-Net requires further post-processing steps to detect the object instances and has been reported to take approximately $20\,\mathrm{ms}$ in conjunction with watershed post-processing [6]. The Cell-DETR variants take the middle ground with $9.0\,\mathrm{ms}$ and $21.2\,\mathrm{ms}$.

V. DISCUSSION

A. Analysis of the instance segmentation performance

Cell-DETR has some benefits in comparison to state-of-the-art methods, such as Mask R-CNN. The Cell-DETR architecture is comparatively simple and avoids common hand designed components of Mask R-CNNs, such as non-maximum suppression and ROI pooling. This reduces Cell-DETR's reliance on hyperparameters and facilitates end-to-end training with a single combined loss function. In contrast, Mask R-CNNs require additional supervision to train the region proposal network. As a result of these differences, Cell-DETR is easier to implement, has less parameters and is faster than Mask R-CNN for the same segmentation performance.

While Cell-DETR does not rely on explicit region proposals, it does utilise attention maps that highlight the pertinent features in the latent space. The mapping of these is learnt during the end-to-end training. The loss curves of individual prediction tasks are shown in Fig. 10. The classification loss $\mathcal{L}_p$ (blue) converges first, indicating that the network first learns how many objects are present in an image and to which class they belong. The bounding box loss $\mathcal{L}_b$ (red) converges next, with the network learning the approximate location of each object. Finally, the model learns to refine the pixel-wise segmentation maps with the segmentation loss $\mathcal{L}_s$ (green) converging last.

With respect to the specific single-cell measurement application, Cell-DETR offers robust and repeatable instance segmentation of yeast cells in microstructures. The key cell class segmentation performance surpasses the previous state-of-the-art semantic segmentation methods [6], [16] with a cell class Jaccard index of 0.84. Additionally, the proposed technique directly detects individual object instances and classifies the objects robustly (near $100\,\%$ accuracy). The robust instance segmentation performance promises to facilitate cell tracking, increase the experimental information yield and enables Cell-DETR to be employed without human intervention.

B. Limitations, outlook and future potential

The presented models are trained for a specific microfluidic configuration and trap geometry. While they are relatively robust and fulfil their intended purpose, their utility could be broadened by expanding the dataset to include more classes, for example different trap geometries. More generally as an instance segmentation method, Cell-DETR offers a platform for incorporating future advances in attention mechanisms as they are increasingly outperforming convolutional approaches. For example, replacing the convolutional elements in the backbone and segmentation head with axial-attention [26] may lead to further improved performance. Currently, Cell-DETR achieves state-of-the-art performance and as an instance segmentation method is generally suitable for and readily adaptable to a wide range of biomedical imaging applications.

The presented Cell-DETR methods can be harnessed for high-content quantitative single-cell TLFM. Cell-DETR, Mask R-CNN and U-Net achieve runtimes orders of magnitudes faster than the previous state-of-the-art trapped yeast method (DISCO [16]). These runtimes coupled with Cell-DETRs robust instance segmentation make both online monitoring and closed-loop optimal experimental design of typical experiments with approximately 1000 traps feasible. Harnessing this potential promises to provide increased experimental information yields and greater biological insights in the future.

VI. CONCLUSION

In summary, we present Cell-DETR, an attention-based transformer method for direct instance segmentation and showcase it on a typical application. To the best of our knowledge, this is the first application of detection transformers on biomedical data. The proposed method has fewer parameters and is 30% faster while matching the segmentation performance of a state-of-the-art Mask R-CNN. A simpler Cell-

DETR variant exhibits slightly lesser segmentation performance ($\mathcal{J}_c = 0.83$ instead of 0.84) while requiring $1/3$rd of a Mask R-CNN's runtime. As a general instance segmentation model, Cell-DETR achieves state-of-the-art performance and is deemed suitable and readily adaptable for a range of biomedical imaging applications.

Showcased on a typical systems or synthetic biology application, the proposed Cell-DETR robustly detects each cell instance and directly provides instance-wise segmentation maps suitable for cell morphology and fluorescence measurements. In comparison to the previous semantic segmentation trapped yeast baselines, Cell-DETR provides better segmentation performance with a cell class Jaccard index $\mathcal{J}_c = 0.84$ while additionally detecting each individual cell instance and maintaining comparable runtimes. This promises to reduce measurement uncertainty, facilitate cell tracking efficacy and increase the experimental data yield in future applications. The resulting runtimes and accurate instance segmentation make future online monitoring feasible, for example for closed-loop optimal experimental control.

Acknowledgements

We thank Christian Wildner for insightful discussions, André O. Françani and Jan Basrawi for contributing to labelling and Markus Baier for aid with the computational setup.

This work was supported by the Landesoffensive für wissenschaftliche Exzellenz as part of the LOEWE Schwerpunkt CompuGene. H.K. acknowledges support from the European Research Council (ERC) with the consolidator grant CONSYN (nr. 773196).

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE ICCV*, 2017, pp. 2961–2969.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE/CVF CVPR*, 2016.

[3] J. Sun, A. Tárnok, and X. Su, "Deep Learning-Based Single-Cell Optical Image Studies," *Cytom. Part A*, vol. 97, no. 3, pp. 226–240, 2020.

[4] M. Leygeber, D. Lindemann, C. C. Sachs, E. Kaganovitch, W. Wiechert, K. Nöh, and D. Kohlheyer, "Analyzing Microbial Population Heterogeneity - Expanding the Toolbox of Microfluidic Single-Cell Cultivations," *J. Mol. Biol.*, 2019.

[5] A. Hofmann, J. Falk, T. Prangemeier, D. Happel, A. Köber, A. Christmann, H. Koeppl, and H. Kolmar, "A tightly regulated and adjustable CRISPR-dCas9 based AND gate in yeast," *Nucleic Acids Res.*, vol. 47, no. 1, pp. 509–520, 2019.

[6] T. Prangemeier, C. Wildner, A. O. Françani, C. Reich, and H. Koeppl, "Multiclass yeast segmentation in microstructured environments with deep learning," *IEEE CIBCB*, 2020.

[7] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *IEEE/CVF CVPR*, 2019, pp. 9404–9413.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv:2005.12872*, 2020.

[9] J.-B. Lugagne, H. Lin, and M. J. Dunlop, "DeLTA: Automated cell segmentation, tracking, and lineage reconstruction using deep learning," *PLoS Comput Biol*, vol. 16, no. 4, 2020.

[10] R. Pepperkok and J. Ellenberg, "High-throughput fluorescence microscopy for systems biology," *Nat. Rev. Mol. Cell Biol.*, p. 690, 2006.

[11] T. Prangemeier, F. X. Lehr, R. M. Schoeman, and H. Koeppl, "Microfluidic platforms for the dynamic characterisation of synthetic circuitry," *Curr. Opin. Biotechnol.*, vol. 63, pp. 167–176, 2020.

[12] D. G. Cabeza, L. Bandiera, E. Balsa-Canto, and F. Menolascina, "Information content analysis reveals desirable aspects of in vivo experiments of a synthetic circuit," in *IEEE CIBCB*, 2019, pp. 1–8.

[13] F.-X. Lehr, M. Hanst, M. Vogel, J. Kremer, H. U. Göringer, B. Suess, and H. Koeppl, "Cell-free prototyping of and-logic gates based on heterogeneous rna activators," *ACS Synth. Biol.*, p. 2163, 2019.

[14] Z. Xie, L. Wroblewska, L. Prochazka, R. Weiss, and Y. Benenson, "Multi-input RNAi-based logic circuit for identification of specific cancer cells," *Science*, vol. 333, pp. 1307–1312, 2011.

[15] W. Si, C. Li, and P. Wei, "Synthetic immunology: T-cell engineering and adoptive immunotherapy," *Synth. Syst. Biotechnol.*, vol. 3, no. 3, pp. 179–185, 2018.

[16] E. Bakker, P. S. Swain, and M. M. Crane, "Morphologically constrained and data informed cell segmentation of budding yeast," *Bioinformatics*, vol. 34, no. 1, pp. 88–96, 2018.

[17] M. M. Crane, I. B. N. Clark, E. Bakker, S. Smith, and P. S. Swain, "A Microfluidic System for Studying Ageing and Dynamic Single-Cell Responses in Budding Yeast," *PLoS One*, vol. 9, p. e100042, 2014.

[18] D. A. Van Valen, T. Kudo, K. M. Lane, D. N. Macklin, N. T. Quach, M. M. DeFelice, I. Maayan, Y. Tanouchi, E. A. Ashley, and M. W. Covert, "Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments," *PLoS Comput Biol*, vol. 12, no. 11, pp. 1–24, 2016.

[19] J. Sauls, J. Schroeder, S. Brown, G. Treut, F. Si, D. Li, J. Wang, and S. Jun, "Mother machine image analysis with MM3," *bioRxiv*, 2019.

[20] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen, "Deep learning for cellular image analysis," *Nat. Methods*, vol. 16, no. 12, p. 1233, 2019.

[21] T. Prangemeier, C. Wildner, M. Hanst, and H. Koeppl, "Maximizing information gain for the characterization of biomolecular circuits," in *Proc. 5th ACM/IEEE NanoCom*, 2018, pp. 1–6.

[22] L. Bandiera, D. Gomez-Cabeza, J. Gilman, E. Balsa-Canto, and F. Menolascina, "Optimally Designed Model Selection for Synthetic Biology," *ACS Synth. Biol.*, 2020.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015, p. 234.

[25] N. Dietler, M. Minder, V. Gligorovski, A. M. Economou, D. A. H. L. Joly, A. Sadeghi, C. H. M. Chan, M. Koziński, M. Weigert, A.-F. Bitbol, and S. J. Rahi, "A convolutional neural network segments yeast microscopy images with high accuracy," *Nat. Commun.*, p. 5723, 2020.

[26] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," *arXiv:2003.07853*, 2020.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF CVPR*, 2016, pp. 770–778.

[28] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML*, 2013, p. 3.

[29] A. Molina, P. Schramowski, and K. Kersting, "Padé activation units: End-to-end learning of flexible activation functions in deep networks," in *ICLR*, 2019.

[30] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *IEEE/CVF CVPR*, 2019, pp. 9308–9316.

[31] H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz, "Pixel-adaptive convolutional neural networks," in *IEEE/CVF CVPR*, 2019, pp. 11 166–11 175.

[32] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[33] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *IEEE/CVF CVPR*, 2019, pp. 658–666.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE ICCV*, 2017, pp. 2980–2988.

[35] K. Chen, J. Wang, J. Pang *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv:1906.07155*, 2019.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.