$\text{ISOO}_{\text{DL}}^{\text{V2}}$ – SEMANTIC INSTANCE SEGMENTATION OF TOUCHING AND OVERLAPPING OBJECTS

Anton Böhm [1] *Maxim Tatarchenko* [1] *Thorsten Falk* [1,2]

[1] Department of Computer Science, [2] BIOSS Centre for Biological Signalling Studies, University of Freiburg, Germany

ABSTRACT

We present $\mathbf{ISOO}_{\mathbf{DL}}^{\mathbf{V2}}$ - a method for semantic instance segmentation of touching and overlapping objects. We introduce a series of design modifications to the prior framework, including a novel mixed 2D-3D segmentation network and a simplified post-processing procedure which enables segmentation of touching objects without relying on object detection. For the case of overlapping objects where detection is required, we upgrade the bounding box parametrization and allow for smaller reference point distances. All these novelties lead to substantial performance improvements and enable the method to deal with a wider range of challenging practical situations. Additionally, our framework can handle object sub-part segmentation. We evaluate our approach on both real-world and synthetically generated biological datasets and report state-of-the-art performance.

Index Terms— Detection, classification, semantic segmentation, instance segmentation, overlapping objects

1. INTRODUCTION

Semantic instance segmentation is one of the central tasks in image understanding. A particularly challenging variant of this task, segmentation of overlapping translucent objects, often occurs in biological images but can also be found in other domains. Many practical applications require not only plain instance segmentation but also the identification of subparts within object instances, e.g. organelles in cells. Existing methods are not able to solve this task for overlapping objects.

In recent years, multiple works approached semantic instance segmentation both on natural and biological images [1–3]. He et al. [4] proposed Mask-RCNN which shows remarkably strong performance on natural images but cannot deal with strongly overlapping objects due to the local non-maximum suppression. In addition, GPU memory requirements get very high when working with images containing many object instances. AffordanceNet [5] adds the possibility to deal with sub-part segmentation but suffers from the same fundamental limitations as Mask-RCNN.

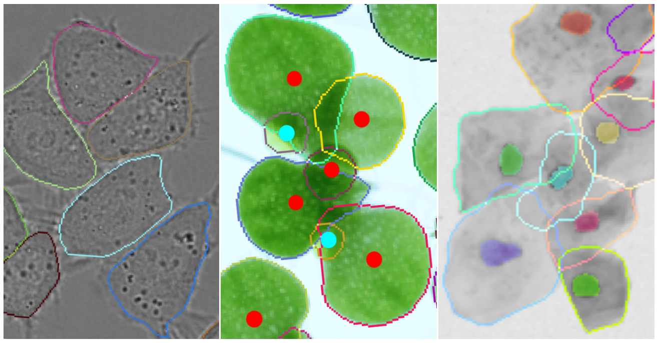

Fig. 1. Our method segments touching (left) and overlapping (middle) objects, at the same time assigning class labels to them (circles of different colors). It can also perform subpart segmentation (right).

In this work we aim at overcoming these restrictions. Our method is based on the $\mathbf{ISOO}_{\mathbf{DL}}$ architecture [6] which includes two separate parts: detection and segmentation. Outputs of both parts are merged in a post-processing step resulting in binary object segmentation masks. The basic idea of $\mathbf{ISOO}_{\mathbf{DL}}$ is to shear the object segmentation masks. This transformation converts them from 2D to 3D and makes them spatially separable. The network predicts such sheared 3D masks. To go back to the original 2D label space, we used the predicted object detections. With $\mathbf{ISOO}_{\mathbf{DL}}^{\mathbf{V2}}$ we pick up the general idea and introduce multiple technical modifications which substantially improve its performance. 1) We allow for more flexible reference point selection in the object detection module. 2) We improve the separation of objects with a small distance between the reference points. 3) We introduce a 2D-3D segmentation network that integrates prior knowledge on valid configurations in all three dimensions of the label space, leading to cleaner segmentation masks. 4) Due to the 3D architecture, our framework supports sub-part segmentation out of the box. 5) The modifications allow us to perform a simplified projection that does not require the detection step and can be easily applied to the case of touching non-overlapping objects. We extensively evaluate the individual building blocks and demonstrate that the final system defines a new state-of-the-art in instance segmentation of overlapping objects on a series of datasets.

We thank the German Federal Ministry for Economic Affairs and Energy (FKz. ZF4184101CR5) and the DFG (EXC 294) for funding our research.

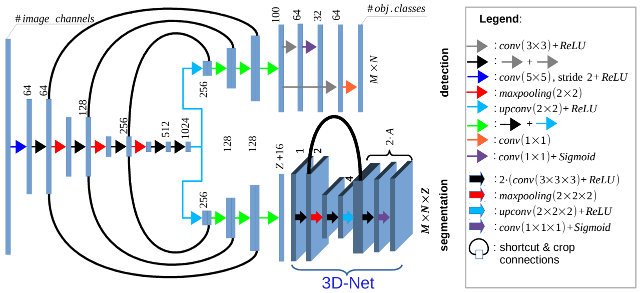

Fig. 2. Hybrid two-headed 2D-3D U-Net architecture.

2. METHOD

We follow the notation of [6]. Every object is represented as a tuple $O_k := (m_k, \mathbf{c}_k, \mathbf{b}_k, \mathbf{p}_k, y_k)$, where m_k is the object mask, $\mathbf{c}_k$ is the center of the bounding box, y_k is the object class label and $\mathbf{b}_k$ is the bounding box size.

In the upcoming sections we present our improvements to the three main blocks of our method: detection (which predicts the object class $\hat{y}_k$ and the bounding box $\hat{\mathbf{b}}_k$), segmentation (which predicts the sheared object masks $(\hat{m}'^x, \hat{m}'^y)$) and post-processing (which merges the last two into the final instance prediction $\hat{O}_k$). The overview of the network is shown in Fig. 2.

2.1. Object bounding box prediction

Reference point location In **ISOO$_{\mathbf{DL}}$** we assumed that the reference points $\mathbf{p}_k$ roughly coincide with the bounding box centers $\mathbf{c}_k$ of objects. Practically, it is often more convenient to instead put the reference points to some unique and representative object locations, e.g. centers of cell nuclei, which can be easily detected but may be located far away from the geometric bounding box centers (Fig. 3a).

Encoding arbitrary reference point locations requires a slight parametrization change. Instead of predicting two dimensions $\mathbf{b}_k = (w_k, h_k)^\top \in \mathbb{R}^2$ along the image axes, we predict the bounding box extents $\mathbf{b}_k = (w_k^\ell, w_k^r, h_k^t, h_k^b)^\top \in \mathbb{R}^4$ in all four directions (left, right, top, bottom) with respect to the reference point $\mathbf{p}_k$, see Fig. 3a. Technically, we discretize each of the four extents into 8 bins and formulate the task as a 32-fold classification.

Dynamic disk size adaptation We describe object reference points as disks which must be spatially isolated to separate overlapping objects. The disk radius is thus an important parameter to define. Smaller disks allow better separation of objects whose reference points are very close to each other. However, the prediction of small disks is particularly hard for the network due to the class imbalance.

To alleviate the problem of manual disk radius selection (Fig. 3b), we allow the disks to shrink in the presence of nearby objects' reference points. First, for every reference point $\mathbf{p}_k$ we compute the euclidean distance $d_k = \min_{j \neq k} ||\mathbf{p}_k - \mathbf{p}_j||_2^2$ to the nearest reference point of another object. Then we calculate the new radius in pixels $\hat{r}_k := \min(\frac{d_k}{2} - 2, r_{\max})$, where $r_{\max}$ is a manually chosen maximum radius (here $r_{\max} = 9$px). To re-balance the effect of disks with different radii on the loss, we introduce per-pixel loss weights $w_k := \min(\pi(r_{\max}^2 - \hat{r}_k^2), 10)$. As before, we additionally force object separation by using inter-object weights.

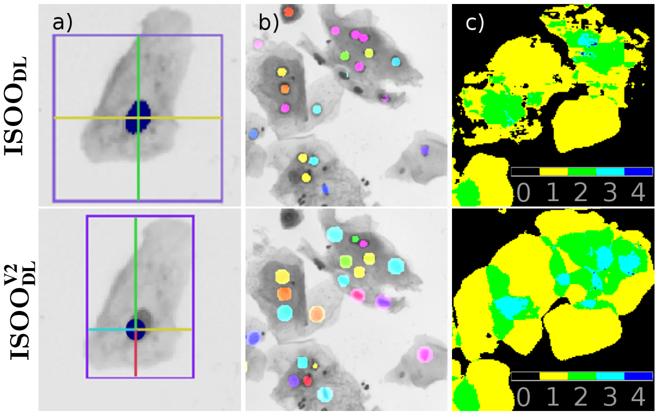

Fig. 3. Comparison of **ISOO$_{\mathbf{DL}}$** and **ISOO$_{\mathbf{DL}}^{\mathbf{V2}}$**. a) We define the extents of the bounding box in 4 directions w.r.t. the object's reference point. b) We introduce adaptive disk sizes to prevent object merging. c) The hybrid 2D-3D network significantly improves the quality of predicted segmentation masks. Colors encode the number of connected components (instance candidates) along Z in $\hat{m}'^x$.

2.2. Sheared object segmentation mask prediction

In **ISOO$_{\mathbf{DL}}$**, we solved the segmentation task by predicting the silhouettes of sheared objects $(\hat{m}'^x, \hat{m}'^y)$ in discretized 3D space using a 2D network. Such formulation has three limitations. First, a 2D network does not optimally exploit correlations in the z-direction of the label space. Second, predicting one pixel-thick connected surfaces in 3D space is challenging; small inaccuracies lead to object fragmentation. Finally, using the channel dimension of the network to represent the spatial Z-coordinate only allows binary (foreground/background) segmentation. Our new architecture (Fig. 2) addresses all these issues.

We modify the sheared object segmentation masks to encode sub-classes: $m'^x, m'^y : \Omega \times \mathbb{Z} \to \{0, \ldots, A\}$, where A is the number of object sub-classes. We also increase their thickness from 1px to 5px. To explicitly model the 3D structure of the output space, we extend the segmentation path with a shallow 3D U-Net with one downsampling step.

In both sheared outputs, the network predicts pseudo-probabilities for A classes and the background. $A = 1$ (cell body) for the problem without sub-classes, and $A = 2$ (cell body and nucleus) for the problem with sub-classes.

We trained the network with input image tiles of shape 508×508px, leading to an output shape of 324×324px for the detection path. To avoid unnecessary computations and save memory, we only use the valid part of the convolution outputs and crop the input to the 3D sub-network so that we obtain an output shape of $144 \times 144 \times Z$ voxels. The Z-range can be adapted to the dataset complexity: higher depth allows a wider range of shearing angles and thus a better separation of objects at the cost of longer training. The exact values of Z are provided in the corresponding paragraphs of Sec. 3.

2.3. Post-processing

In $\mathbf{ISOO_{DL}}$ we compute the final object masks by applying the inverse transformation, part of which is the weighted projection onto the central object planes. The detection results are required to extract the corresponding objects from the sheared masks. Here we propose a simpler and more robust projection scheme that works even without using the detection outputs in the case of touching non-overlapping objects.

Consider a foreground/background segmentation task of **non-overlapping** objects. The improved segmentation path allows to extract the predicted 3D planes in both sheared segmentation outputs via connected component labeling. To identify the corresponding objects in two sheared outputs, we consider all possible pairs of x-sheared and y-sheared connected components. When projected to the x-y-plane, binarized and multiplied, only masks belonging to the same object produce a non-zero overlap.

When objects in the original image do **overlap**, the described procedure generates some false positives – overlapping regions. However, the predicted bounding boxes allow to easily identify and discard those. We assign the bounding boxes to their corresponding extracted connected components using the Hungarian algorithm [7]; the costs are set to $1 - \text{IoU}$ between the bounding box and the candidate segment. Connected components which have a corresponding bounding box are accepted as final object instance predictions.

In regions with many overlapping objects this simple approach fails due to merging of very closely located sheared object masks. Therefore, we fall back to the original backprojection based on reference points, if the IoU between a segment and any bounding box drops below 0.3. If the final IoU value between the segmentation mask and the best-fitting bounding box is still smaller than 0.3, we simply replace the segmentation mask by the filled bounding box.

With the same procedure, we can also identify **sub-parts**. The only additional step is to make sure that only the sub-parts belonging to the same object are matched after the connected component extraction. This is achieved by multiplying the sheared sub-part segmentation masks with the sheared binary object segmentation masks.

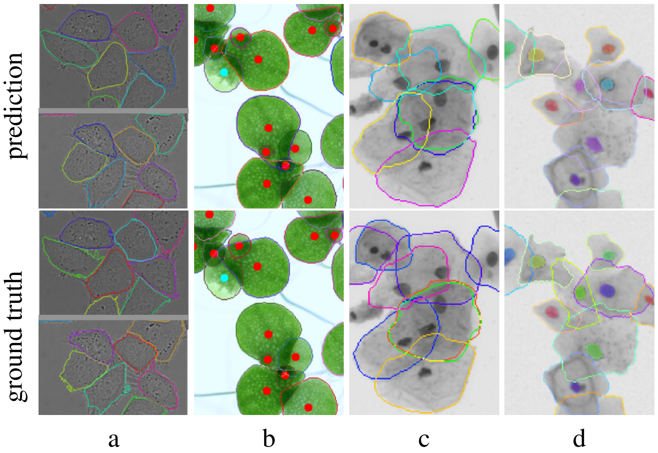

Fig. 4. Qualitative results on different datasets. a) **DIC-Hela**, b) **Duckweed** (colors represent different object classes), c) **OSC-ISBI**, d) **OSC-ISBI-S** (object contours and the corresponding sub-parts have the same color).

3. EXPERIMENTS

3.1. Datasets

We test our method on 4 datasets. We use the **"DIC-Hela"** dataset from the ISBI Cell Tracking Challenge [8] for instance segmentation of touching non-overlapping objects. We downscale the images by a factor of two and hold out 50% of the available training set for validation. Experiments with overlapping objects are performed on the same datasets as in [6]: **"Duckweed"** and **"OSC-ISBI"** [9]. The task in the **"Duckweed"** dataset is to segment all leaflets of the duckweed rosettes and classify each object as either healthy or showing chlorosis. This dataset is particularly challenging because some leaflets are completely covered and there are only very few small chlorosis areas.

The **"OSC-ISBI"** dataset does not have sub-part annotations for the test set, therefore it cannot be used to evaluate the sub-part segmentation performance. Instead, we make use of an existing dataset of synthetically generated cells [10]. We call this dataset **"OSC-ISBI-S"**. We do not use all 900 available test images, but restrict ourselves to the most complex cases (images with at least 10 objects).

As in [11] and [6], we evaluate the performance with dice coefficient **DC**, pixel-based true positive **TPp** and false positive **FPp** rates as well as with object-based false negative **FPp** scores.

3.2. Results

In the first three experiments we analyze the performance of individual modules of our pipeline.

1. Detection-only network. We compare the detection performance with and without dynamic disk size adapta-

	DC	FNo	TPp	FPp
Phoulady *et al.* [12]	.831±.079	.408±.163	.927±.098	.003±.002
Ramalho *et al.* [13]	.856±.078	.501±.180	.899±.113	.002±.001
Lee *et al.* [14]	.879±.087	.434±.168	.877±.123	.001±.001
ISOO$_{DL}$ [6]	.863±.074	.370±.141	.895±.107	.001±.001
ours	.895±.079	.290±.151	.901±.108	.001±.001

Table 1. Results on the **”OSC-ISBI”** dataset without sub-part segmentation (Mean ± SD).

tion (see Fig. 3b) on the **”OSC-ISBI”** dataset (segmentation branch not used). We compare the predicted bounding boxes with the bounding boxes of the ground truth segmentation masks and report an **FNo** of .361±.113 without and .285±.111 with disk size adaptation.

2. Segmentation-only network. We compare the segmentation branch of the original $\mathbf{ISOO_{DL}}$ 2D network with the proposed hybrid 2D-3D network on the **”OSC-ISBI”** dataset. For this, we delete the detection branch of the $\mathbf{ISOO_{DL}^{V2}}$ network and compute the IoU between the sheared ground truth and the predicted segmentation masks $\hat{m}'^{x}$ and $\hat{m}'^{y}$ after binarization (binarization threshold is set to 0.5). We trained the $\mathbf{ISOO_{DL}}$ network on the thickened (5px) masks for fair comparison. We report an IoU of 0.29 and 0.33 without and with the stacked 3D network respectively. However, IoU alone does not account for object fragmentation which is very important for our current projection method. To analyze this property, we also calculate the ratio between the number of predicted and the ground truth connected components. We report 0.34 without and 0.5 with the 3D network respectively. The low ratio for the 2D network demonstrates that it leads to severe over-segmentation (Fig. 3c).

3. Projection. We test the performance of the projection module on **”DIC-HeLa”** which does not contain overlapping objects and thus allows extracting predictions solely from the segmentation masks. When using the new projection without detections, we get **DC** and **FNo** scores of .911±.050 and .083±.061 respectively versus .889±.067 and .113±.099 for $\mathbf{ISOO_{DL}}$. Due to the absence of overlapping objects, such segmentation-only projection cannot produce false positives mentioned in 2.3 and thus using detections is not required. However, to show that our new way of integrating the segmentation and detection results is more robust than the one from $\mathbf{ISOO_{DL}}$, we also evaluate the full projection and get a **DC** of .903±.057 and an **FNo** of .093±.074. A slight drop w.r.t. the segmentation-only case is due to some false negatives produced by the detection module.

4. Instance segmentation of overlapping objects without sub-parts. We test the performance on **”OSC-ISBI”** (Fig. 4.3). As in [6], we use the centers of mass of the cell nuclei as reference points. Since the **”OSC-ISBI”** dataset does not provide a mapping between the nuclei and their corresponding cells, we assigned them as follows. We assume that the nucleus is located in the cell center. First, we separate merged nuclei using the Watershed transform, then we assign the nucleus segments to cells using the Hungarian method based on their distances from the cells' centers of mass. Unmatched nucleus segments are assigned to the closest cell (according to the described metric). The dataset contains many densely overlapping objects, therefore we decided to downsample the images with factor 0.5 and set $Z = 100$ to get a higher shearing angle. We initialize the network with the weights from 3.2.2 and train it for another 250K iterations. We compare the performance to $\mathbf{ISOO_{DL}}$ and report a new state-of-the-art in all metrics except for **TPp** (Tab. 1).

	DC	FNo	TPp	FPp
ISOO$_{DL}$ [6] (*cl*)	.913±.077	.314±.202	.937±.068	.000±.000
ours (*cl*)	.914±.071	.296±.187	.930±.061	.000±.000
ISOO$_{DL}$ [6] (*w/o cl*)	.929±.057	.129±.067	.939±.066	.000±.000
ours (*w/o cl*)	.945±.051	.103±.062	.953±.063	.000±.000

Table 2. Results on the **”Duckweed”** dataset (Mean ± SD).

To show that our method generalizes to other types of data, we apply it to **”Duckweed”** (Fig. 4b). Similar to [6], we use a lower resolution of $Z = 64$. The performance is comparable to [6] when computing the scores for each class individually and then averaging over them (Tab. 2,*”cl”*). When evaluating the performance without class distinction, our approach sets the new state of the art (Tab. 2,*”w/o cl”*).

5. Instance segmentation of overlapping objects with sub-parts. For this experiment we extend our segmentation branch output by two extra channels, initialize it with the weights from 3.2.4 and train it for 250K iterations. The evaluation metrics stay the same with a small extension: we compute the quality of ”good” segmentation masks (DC>0.7) not only for the cell body, but also for the nuclei (written in scopes). We get an **FNo** of .210±.194, a **DC** of .924(.848)±.069(.173), a **TPp** of .906(.842)±.102(.198) and an **FPp** of .000(.000)±.001(.000). The segmentation results for the nucleus are slightly worse than for the cell body for several reasons. First, the class imbalance problem becomes more relevant for small object parts, especially in 3D space. Second, since we also want to separate the overlapping subparts, we need to multiply the sub-part predictions with the differently sheared cell body predictions in the projection step; if any of the segmentations fails, the overall performance will be bad. Finally, assignment of nuclei to their corresponding cells is hard in high-density regions.

4. DISCUSSION

In this work we introduced a series of improvements to the original $\mathbf{ISOO_{DL}}$ leading to significant performance increases. Although theoretically $\mathbf{ISOO_{DL}^{V2}}$ could solve any instance segmentation problem, in practice the available GPU memory limits the maximum object size. A possible solution is to replace the memory-intensive dense 3D decoder with a recent quadtree/octree-based network ([15], [16]).

5. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[2] Jonas Uhrig, Marius Cordts, Uwe Franke, and Thomas Brox, "Pixel-level encoding and depth layering for instance-level semantic labeling," in *German Conference on Pattern Recognition*, 2016.

[3] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei, "Fully convolutional instance-aware semantic segmentation," *arXiv preprint arXiv:1611.07709*, 2016.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask R-CNN," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[5] Thanh-Toan Do, Anh Nguyen, and Ian Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *International Conference on Robotics and Automation (ICRA)*, 2018.

[6] Anton Böhm, Annekathrin Ücker, Tim Jäger, Olaf Ronneberger, and Thorsten Falk, "Isoo dl: Instance segmentation of overlapping biological objects using deep learning," in *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th*, 2018.

[7] Harold W Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, 1955.

[8] V. Ulman, M. Maška, K. E. G. Magnusson, O. Ronneberger, C. Haubold, and et al., "An objective comparison of cell-tracking algorithms," *Nature Methods*, 2017.

[9] "The second segmentation of overlapping cervical cells from extended depth of field cytology image challenge," `https://cs.adelaide.edu.au/~zhi/isbi15_challenge/`, The Second Segmentation of Overlapping Cervical Cells from Extended Depth of Field Cytology Image Challenge is held under the auspices of the IEEE International Symposium on Biomedical Imaging (ISBI 2015) held in New York, USA on April 16th - 19th, 2015.

[10] "The first segmentation of overlapping cervical cells from extended depth of field cytology image challenge," `https://cs.adelaide.edu.au/~zhi/isbi15_challenge/`, The First Segmentation of Overlapping Cervical Cells from Extended Depth of Field Cytology Image Challenge is held under the auspices of the IEEE International Symposium on Biomedical Imaging (ISBI 2014) held in Beijing, China on April 28th - May 2nd, 2013.

[11] Zhi Lu, Gustavo Carneiro, Andrew P Bradley, Daniela Ushizima, Masoud S Nosrati, Andrea GC Bianchi, Claudia M Carneiro, and Ghassan Hamarneh, "Evaluation of three algorithms for the segmentation of overlapping cervical cells," *IEEE journal of biomedical and health informatics*, 2017.

[12] Hady Ahmady Phoulady, Dmitry B Goldgof, Lawrence O Hall, and Peter R Mouton, "An approach for overlapping cell segmentation in multi-layer cervical cell volumes," *The Second Overlapping Cervical Cytology Image Segmentation Challenge-IEEE ISBI*, 2015.

[13] Geraldo LB Ramalho, Daniel S Ferreira, Andrea GC Bianchi, Claudia M Carneiro, Fátima NS Medeiros, and Daniela M Ushizima, "Cell reconstruction under voronoi and enclosing ellipses from 3d microscopy," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2015.

[14] Hansang Lee and Junmo Kim, "Segmentation of overlapping cervical cells in microscopic images with superpixel partitioning and cell-wise contour refinement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[15] Gernot Riegler, Ali Osman Ulusoy, Horst Bischof, and Andreas Geiger, "Octnetfusion: Learning depth fusion from data," in *Proceedings of the International Conference on 3D Vision*, 2017.

[16] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs," in *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017.