Automated Generation of Instance Segmentation Labels for Traffic Surveillance Models

D. Scholte[2][a], T. T. G. Urselmann[1][b], M. H. Zwemer[1,2][c], E. Bondarev[1] and P. H. N. de With[1][d]

[1]*Department of Electrical Engineering, Eindhoven University, Eindhoven, The Netherlands*

[2]*ViNotion BV, Eindhoven, The Netherlands*

Keywords: Instance Segmentation, Object Detection, Real-Time Processing, Computer Vision, Traffic Surveillance.

Abstract: This paper focuses on instance segmentation and object detection for real-time traffic surveillance applications. Although instance segmentation is currently a hot topic in literature, no suitable dataset for traffic surveillance applications is publicly available and limited work is available with real-time performance. A custom proprietary dataset is available for training, but it contains only bounding-box annotations and lacks segmentation annotations. The paper explores methods for automated generation of instance segmentation labels for custom datasets that can be utilized to finetune state-of-the-art segmentation models to specific application domains. Real-time performance is obtained by adopting the recent YOLACT instance segmentation with the YOLOv7 backbone. Nevertheless, it requires modification of the loss function and an implementation of ground-truth matching to overcome handling imperfect instance labels in custom datasets. Experiments show that it is possible to achieve a high instance segmentation performance using a semi-automatically generated dataset, especially when using the Segment Anything Model for generating the labels.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Automated traffic surveillance systems support a range of tasks involving congestion and accident observation or crowd management analysis. In these systems, cameras are generally used to find the trajectories of all relevant traffic participants in a scene. In order to analyse the behaviour of traffic participants, it is vital to accurately localize and follow all objects over time. Typical (real-time) techniques for object localization use 2D bounding boxes to represent the object location. However, an instance segmentation of an object provides more accurate localization, especially for large elongated objects such as a truck as depicted in Figure 1. Only bounding boxes do not provide the insight on the real top-view central point of an actor, instance segmentation is a more refined technique that enables computation of the central point especially if the camera parameters are known (Zwemer. et al., 2022).

The focus of this work is on extending an object detection model with instance segmentation that can be utilized for real-time traffic analysis. It is important to run instance segmentation in parallel to bounding box estimation, since a bounding box covers the complete traffic participant (even if they are partially occluded), while the instance segmentation is only available for the visible parts of the object. Instance segmentation models are a hot topic in literature (Sharma et al., 2022) and although the amount of models available is large and ever-growing, the majority of these models are computationally complex. Another challenge is the lack of instance segmentation datasets for training and evaluation, since current state-of-the-art datasets are not specifically aimed at traffic surveillance. The creation of ground-truth

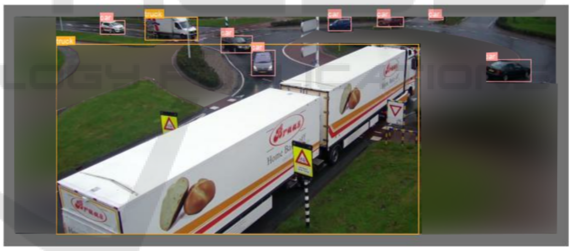


Figure 1: A scene containing a large elongated object that has a lot of background within the bounding box. Blurred for privacy reasons.

[a] https://orcid.org/0009-0004-9182-2911
[b] https://orcid.org/0000-0002-2209-7216
[c] https://orcid.org/0000-0003-0835-202X
[d] https://orcid.org/0000-0002-7639-7716

Scholte, D., Urselmann, T., Zwemer, M., Bondarev, E. and de With, P.
Automated Generation of Instance Segmentation Labels for Traffic Surveillance Models.
DOI: 10.5220/0012319500003660
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 350-358
ISBN: 978-989-758-679-8; ISSN: 2184-4321
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

for instance segmentation is cumbersome and time-consuming. A proprietary dataset containing traffic surveillance images annotated with bounding boxes only is available for our experimentation.

The problem statement addressed in this paper is to explore a suitable instance segmentation model that can be finetuned on the proprietary dataset, without the need for manual annotation of the data. To this end, we experiment with the YOLACT-YOLOv7 model that is able to perform object detection and instance segmentation in real-time. However, this model is not trainable without instance segmentation ground truth. Therefore, this paper investigates the following research questions:

- What segmentation model can be utilized best for real-time object detection and instance segmentation in traffic surveillance applications?
- To what extent can this model be optimized such that the best performance is achieved, while still achieving a real-time performance?
- What solutions can be applied for the absence of ground-truth data for instance segmentation in the proprietary dataset annotated with bounding boxes only?

The remainder of the paper is structured as follows. A literature review of state-of-the-art models is given in Section 2. The methodology of proposed strategies is presented in Section 3. Section 4 discusses the experimental setup and results. Lastly, Section 5 summarizes and concludes this research.

2 RELATED WORK

This section presents a brief overview of the large variety of instance segmentation models. Recent models can be divided into several categories. Instance segmentation models are typically trained fully-supervised with ground-truth segmentation annotations. These models can be categorized into two-stage and single-stage.

Two-Stage Models. Similarly to object detection, two-stage models, such as Mask R-CNN (He et al., 2017), first create object proposals at the first stage and refine these proposals at the second stage. Single-stage approaches create proposals and perform the refinement in one shot and are typically more computationally efficient than two-stage approaches.

Single-Stage Models. Single stage approaches are SOLO(v2) (Wang et al., 2020a; Wang et al., 2020b), YOLACT (Bolya et al., 2019), and more recent also BlendMask (Chen et al., 2020), which is an extension to YOLACT. Within YOLACT and BlendMask, different activation maps are combined to dictate the instances. YOLACT uses coefficients to determine the combination of these activation maps, on the other hand, BlendMask employs attention maps based on activation maps, both being computationally efficient and resulting in final instances.

Recently, YOLOv7 (Wang et al., 2023) has shown to be an effective and particularly fast model that is suited for real–time object detection. It has been adapted to serve as a backbone for YOLACT (Munawar and Hussain, 2023). This model achieves competitive performance compared to other single-stage instance segmentation models but is compact and has a low-latency inference.

Transformer-Based Models. Besides these two categories of convolutional neural networks, transformer models have recently shown their capabilities to achieve a high detection and segmentation performance. The recent Mask2Former (Cheng et al., 2021) consists of a backbone, a pixel encoder and a transformer decoder. Its main feature is masked attention, which is a variant of cross attention but constrained on mask query prediction. Another transformer model with high performance is the Segment Anything Model (SAM) (Kirillov et al., 2023). It utilizes an image encoder based on the Vision Transformer algorithm (Dosovitskiy et al., 2020) to produce image embeddings. Prompts are then used to determine the embeddings of interest. These prompts can be divided into two categories: sparse prompts (being points, bounding boxes, or text) or dense prompts (being masks). The mask decoder processes the prompts together with the image embeddings to create a high-quality mask. This model is computationally heavy to be used for edge devices, but can be utilized to generate segmentation labels to build a training dataset.

Box-Supervised Models. In instance segmentation literature, there is a paradigm shift to box-supervised instance segmentation models because of the limited amount of ground-truth instance segmentation. In these techniques, only bounding-box annotations are used for supervision during training. The recent BoxLevelset (Li et al., 2022a) builds upon the SOLOv2 model and utilizes an instance-aware decoder that is improved by a level-set evolution step within training. This step includes the Chan-Vese energy function (Getreuer, 2012) to evaluate the segmentation performance based on the bounding box ground truth. The Box2Mask (Li et al., 2022b) acts as an improvement of BoxLevelset by introducing a Local Consistency Module (LCM) that exploits local pixel consistencies and this model has been implemented for both the SOLOv2 and the MaskFormer

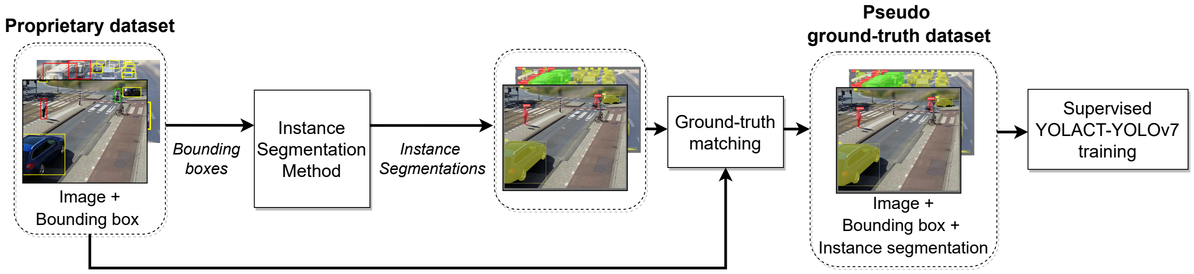


Figure 2: Training pipeline with pseudo ground-truth data. The instance segmentation method utilized for generating the pseudo ground-truth data is either a box-supervised model that is finetuned using the proprietary data or a pre-trained model. The pseudo ground-truth data is used to train the final YOLACT-YOLOv7 model.

models (Li et al., 2022b). These box-supervised models, especially Box2Mask, show promising results and the performance gap between fully-supervised and box-supervised instance segmentation is decreasing significantly. These models could be finetuned on our proprietary dataset, such that the performance is optimized, and be used to generate segmentation labels for the whole dataset afterwards.

3 METHOD

This section presents the proposed methodology for training the YOLACT-YOLOv7 model on the proprietary dataset, without the requirement to manually create instance segmentation ground-truth for the dataset. To that end, Section 3.1 proposes two novel semi-automated approaches to generate instance segmentation ground-truth for the proprietary dataset, these methods are based on existing instance segmentation models and on box-supervised models that are first finetuned on the proprietary dataset, respectively. Section 3.2 continues by proposing an adjustment of the loss function of the YOLACT-YOLOv7 model, such that it is able to handle datasets that are annotated with segmentation labels only partially.

3.1 Generating Pseudo Ground-Truth

This section investigates the use of models that can generate pseudo ground-truth data for instance segmentation semi-automatically. The generation of pseudo ground-truth is proposed in two different approaches. First, the existing pre-trained models are utilized (see Section 3.1.1). Second, box-supervised models are finetuned on the proprietary dataset (see Section 3.1.2) and then utilized to create pseudo ground-truth labels. Let us discuss both methods in detail.

3.1.1 Pre-Trained Generation of Instances

Figure 2 depicts the automated processing pipeline that includes multiple steps to create the ground truth for the proprietary dataset and finally train the YOLACT-YOLOv7 model on this data. First, the entire proprietary dataset containing bounding boxes is processed by a pre-trained model to generate instance segmentations for all objects in the dataset. Next, these instance segmentations need to be matched with the bounding boxes from the proprietary dataset to create the pseudo ground-truth dataset. This matching procedure is now discussed in more detail.

Ground-Truth Matching. The instance segmentations from the model are predicted independently (not related to the bounding-boxes in the pseudo ground-truth). The number of generated instances and the order of the instances can differ from the bounding-box annotations, therefore the segmentations cannot be linked directly to the bounding boxes. For each generated segmentation the Complete Intersection over Union (CIoU) is computed by using the bounding box around the edges of the segmentation mask and each bounding box from the ground-truth. Then, matches are generated based upon the Hungarian matching algorithm (Kuhn, 2012), e.g. the highest matching pair based on the CIoU is selected for the pseudo ground-truth dataset and removed from the set of possible matches. This is repeated until the set of possible matches is empty or the CIoU values do not exceed a threshold of 0.30. It is possible that not all bounding boxes are matched with an instance segmentation because of possible missed or false segmentations generated by the instance segmentation model. Therefore, the ground-truth bounding boxes that are not matched to any instance segmentation are also added to the pseudo ground-truth dataset (without segmentation). To enable training the YOLACT-YOLOv7 model with data that does not contain segmentation ground-truth for all bounding box annotations (partially annotated data), a novel loss function is pro-

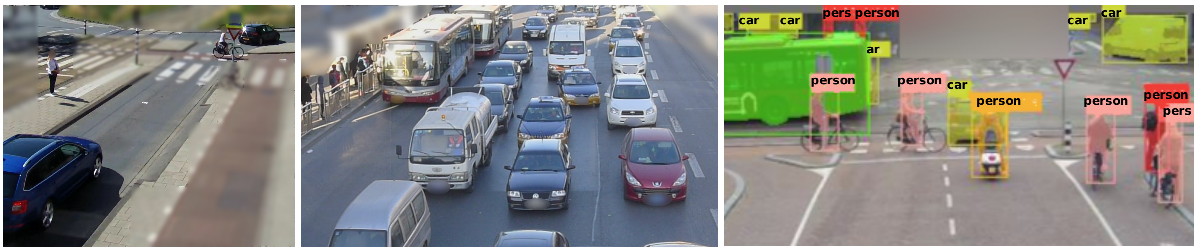


Figure 3: Example images of the proprietary dataset used for this research. The image on the right-hand side is manually annotated with instance segmentations. Blurred for privacy reasons.

posed in Section 3.2.

Training of the YOLACT-YOLOv7 Model. The last step of Figure 2 is the supervised training of the YOLACT-YOLOv7 model. The created pseudo ground-truth dataset that is generated by the ground-truth matching is used to train the YOLACT-YOLOv7 model, including the proposed loss function to handle the imperfections remaining in the pseudo ground-truth dataset.

3.1.2 Box-Supervised Generation 0f Instances

Pre-trained models for instance segmentation are not adapted to the traffic surveillance domain. Ideally, these models are adapted for traffic surveillance in order to improve the segmentation accuracy, this can be achieved by finetuning box-supervised instance-segmentation models on the proprietary dataset. Finetuning a model without segmentations is impossible for fully-supervised models. Nevertheless, box-supervised instance segmentation models can be fine-tuned using solely bounding box annotations. Thus, it is possible to fine-tune these models using the proprietary dataset such that a higher segmentation performance is achieved. This approach is shown as the second method of Figure 2. After finetuning, these models can generate annotations automatically for the proprietary dataset. The ground-truth matchting, the adapation of the loss function and the training of YOLACT-YOLOv7 model in this approach are similar to the first method.

3.2 Adaptation Loss-Function

The original YOLACT-YOLOv7 model can only be trained with datasets that contain both the bounding box and segmentation ground-truth for each object. The instance-segmentation annotations are currently not available in our proprietary traffic surveillance dataset. Therefore, we propose to adjust the loss function such that the model can be trained with a dataset that is only partly annotated with instance segmentations. This is an interesting approach since the proprietary dataset can be combined with generic public datasets that have instance segmentation annotations available. Hence, this may relieve the annotation effort on the proprietary dataset.

In more detail, the YOLACT-YOLOv7 loss consists of four different loss components, e.g. the objectness $\mathcal{L}_{\text{obj}}$, classification $\mathcal{L}_{\text{cls}}$, bounding-box $\mathcal{L}_{\text{box}}$ and mask $\mathcal{L}_{\text{mask}}$ losses. In the proposed loss function, the contribution of the mask loss is dependent on the amount of objects that include instance segmentation data within a batch, such that the limited amount of annotations are automatically weighted proportionally to the other loss terms. Therefore, the proposed loss function is as follows:

$$\mathcal{L} = \lambda_{obj}\mathcal{L}_{\text{obj}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \frac{\lambda_{\text{msk}}\alpha_{\text{msk}}}{\sum_{i=0}^{N}\frac{gt_{\text{msk,i}}}{gt_{\text{all,i}}}}\mathcal{L}_{\text{msk}}, \qquad (1)$$

where λ_{obj}, λ_{cls}, λ_{box} and λ_{msk} are scalar weights for the respective loss functions, N is the batch size, $gt_{\text{mask,i}}$ is the number of ground-truth masks within image i, $gt_{\text{all,i}}$ is the total number of ground-truth annotations within an image i, and α_{mask} is a hyperparameter that denotes the fraction of objects in the dataset that have instance segmentation ground-truth available (i.e. if 70% of the annotation data contain both a bounding box and a segmentation, then α_{mask} becomes 0.7).

4 EXPERIMENTS

In Section 4.1, cross-validation is conducted among various instance segmentation models applied to the proprietary validation set. The second experiment in Section 4.2 uses partially annotated data to measure the impact of training the YOLACT-YOLOv7 model with only a fraction of segmentation instance labels and a full set of bounding boxes. Thereafter, the creation of instance segmentation ground-truth for the proprietary dataset is investigated in Section 4.3. In the final experiment in Section 4.4, the generated instance segmentation labels on the proprietary dataset

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 1: Cross-validation on state-of-the-art instance segmentation models using the manually annotated proprietary validation set. All *mAP* results are shown in percentage.

Model	**Backbone**	**Box** $\mathbf{mAP}_{0.50}$	**Box** $\mathbf{mAP}_{0.50-0.95}$	**Mask** $\mathbf{mAP}_{0.50}$	**Mask** $\mathbf{mAP}_{0.50-0.95}$	**Inf. time** **[ms]**
Mask R-CNN	ResNet50	75.4	58.9	74.6	56.5	198.75
YOLACT	ResNet101	67.3	45.3	64.0	45.7	121.99
BlendMask	ResNet101	**79.2**	**65.3**	78.5	62.7	182.53
SOLOv2	ResNet101	77.5	65.0	75.4	58.0	292.43
QueryInst	ResNet101	76.6	63.2	77.6	59.6	334.27
YOLACT	YOLOv7	78.4	64.6	76.6	57.6	**21.2**
Mask2Former	Swin-S	76.6	62.1	**80.1**	**62.9**	530.93
BoxLevelset	ResNet50	65.5	44.9	64.7	43.7	258.75
Box2Mask	ResNet101	78.2	61.1	79.1	57.8	799.36

are deployed to train the YOLACT-YOLOv7 model, and its performance is measured.

Experimental Setup. All experiments on the YOLACT-YOLOv7 model keep the original parameters. The values of λ_{obj}, λ_{cls} and λ_{box} in Equation (1) are set to 0.7, 0.3, and 0.05, respectively. The number of prototypes is set to 32.

Traffic Surveillance Dataset. The proprietary dataset includes 130k images of traffic surveillance scenes. This dataset includes a variety of scenes such as complex intersections, crowded pedestrian places, and busy highways. Example images are shown in Figure 3. The dataset contains bounding-box annotations only, which is a major limitation for training an instance segmentation model. The annotated bounding boxes contain the entire body (e.g. the boxes cover occluded body parts). The four relevant classes are Person, Car, Bus, and Truck.

Proprietary Validation Dataset. Since the dataset contains bounding boxes only, the validation of segmentation models is impossible. Therefore, 100 images have been manually annotated with instance segmentations (proprietary validation set), containing 1.230 persons, 396 cars, 23 trucks and 21 busses. The proprietary validation set is representative for the proprietary dataset and is utilized for validation in all experiments. An example of an annotated image can be seen on the right-hand side in Figure 3.

Evaluation metrics. The metrics used for evaluation of the instance segmentation and detection performance are based on the COCO protocols. The Average Precision (AP) is calculated by using an Intersection over Union (IoU), either an IoU of 0.5 ($\mathrm{AP}_{0.5}$) or averaged over the range [0.5:0.05:0.95] ($\mathrm{AP}_{0.50-0.95}$). The AP is calculated for the bounding boxes and the segmentation masks separately. The precision and recall for segmentation masks are calculated per pixel between the ground truth and the predicted mask. The mean Average Precision (mAP) is the average value of the AP over all classes. The inference-time measurements are defined by the average time that the model requires to process the proprietary validation set.

4.1 Model Cross-Validation

The performance of existing (pre-trained) models on our proprietary surveillance dataset is investigated in the first experiment. The objective is to find a model that has a low latency while having a high segmentation performance. The models are evaluated by cross-validation on the proprietary validation set.

The results are depicted in Table 1. The BoxLevelset model has lower performance compared to the other models. However, this model is box-supervised during training and does not require instance segmentation ground-truth during training. The Box2Mask model, that is also box-supervised, achieves higher performance and even competes with fully-supervised instance segmentation models in terms of detection and segmentation performance. Another interesting result is that there are major differences between the ResNet101-based and YOLOv7-based YOLACT models. With respect to performance, the YOLOv7 backbone is more efficient and results in significantly higher performance. Moreover, the YOLACT-YOLOv7 implementation has a significantly lower inference time with respect to all other models, while having sufficiently accurate detection and segmentation performance. Therefore, the YOLACT-YOLOv7 implementation is selected for further experiments.

4.2 Training with Fraction of Instances

This experiment helps to determine whether it is necessary to provide both bounding boxes and instance segmentation annotations for all data in an instance segmentation dataset, or whether it is sufficient to provide only a subset of the data with instance segmentation annotations. The latter would imply a reduction in the required (manual) annotation effort or training

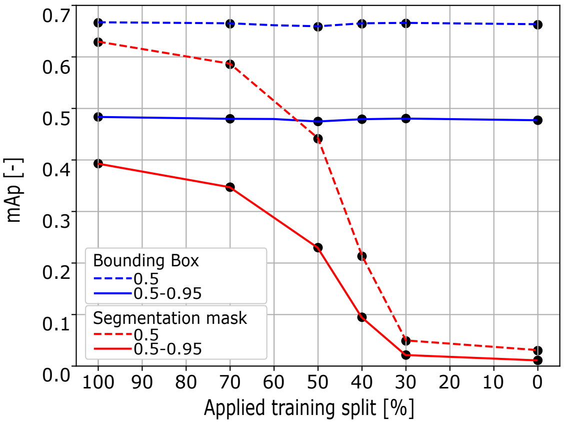


Figure 4: Instance segmentation performance scores for training with a limited amount of instance segmentations. The $mAP_{0.5}$ is depicted as dashed lines and the $mAP_{0.5-0.95}$ is depicted as solid lines. The segmentation performance starts to drop heavily below a split of 70%, while the box performance remain high.

with a imperfect pseudo ground-truth dataset.

This experiment evaluates the effect of YOLACT-YOLOv7 training with only a fraction of the instance labels and is performed with the COCO2017 dataset (Lin et al., 2014). To evaluate the amount of required segmentation ground-truth, splits between data with and without instance labels are used.

The results of the experiment are shown in Figure 4. It can be observed that for each applied split, the bounding-box performance remains high within a small deviation. However, a clear drop in segmentation performance occurs when less than 70% of instance segmentation annotations are available for training. Below a 50% split, the segmentation performance deteriorates significantly.

From these results, it can be concluded that it is possible to achieve decent segmentation performance when not all images in the dataset are annotated with instance segmentations. Up to 90% of the segmentation performance can be achieved for the COCO dataset with only 70% of instance segmentation annotations. Therefore, acceptable results are expected on the proprietary dataset if at least 70% of the dataset is annotated with instance segmentation labels. However, 70% is still deemed too cumbersome to generate for the proprietary dataset manually.

4.3 Creation of Pseudo Ground-Truth

This experiment evaluates generation of instance segmentation for the proprietary dataset. This pseudo ground-truth is generated semi-automatically. The procedure is described in Section 3.1. The validation of the models is performed on the proprietary dataset.

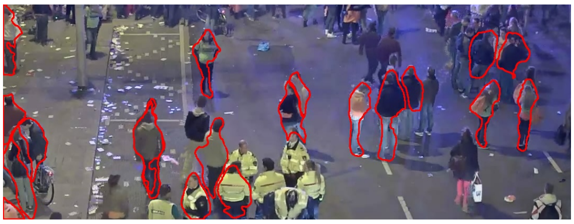


Figure 5: Generated instance labels from the pretrained Box2Mask model. There are major difference between the COCO-dataset and the proprietary dataset, resulting in a lot of missing instance labels. Blurred for privacy reasons.

Besides the bounding box and segmentation metrics, the number of generated segmentation instances that are matched to the ground-truth bounding boxes are measured. The results for the pre-trained models and the fine-tuned box-supervised models are now separately discussed in more detail.

4.3.1 Label Generation Using Pre-Trained Models

In the first approach, we investigate the creation of the instance segmentation labels for the proprietary dataset by utilizing pre-trained models such as the BoxLevelSet, Box2Mask and SAM.

The results are presented in the top three rows in Table 2. It can be observed that the Box2Mask model achieves better performance than the BoxLevelset model. Nevertheless, visual inspection shows that there are missing predictions for objects in the back of the scene and occluded objects depicted in Figure 5. Besides that, SAM has the best segmentation results for creating the pseudo ground-truth labels with a mask mAP score of 84.3% and always generates a segmentation for each bounding-box annotation in the ground truth.

4.3.2 Label Generation Using Fine-Tuned Models

In the second approach, the box-supervised models are fine-tuned on the proprietary dataset. Hence, better performance is expected compared to the models evaluated in the previous experiment.

A limitation at box-supervised models such as Box2Mask and BoxLevelset, is that they are too large in terms of memory consumption to train with the default settings. Therefore, it is chosen to use half of the image resolution, thereby reducing the memory consumption by a factor of four [1]. The loss weight-

[1] For indication, the regular training of BoxLevelset required 8x V100 GPU with 32 GB memory each, whereas for this research a GPU setup of 3x 3090 GPU with 24 GB memory was available.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 2: Results of the BoxLevelset and Box2Mask model using a pre-trained model on the COCO dataset, and after finetuning on the proprietary dataset. The second half of the table shows models that include finetuning on the proprietary dataset. It should be noted that for SAM, the original resolution of the input image is used and no matching is required since it uses the bounding boxes from the ground truth as input prompts (100% matching score and Box mAP).

Model	**Input resolution**	**Training dataset**	**Ground truth matching**	**Box mAP$_{0.5}$**	**Box mAP$_{0.50-0.95}$**	**Mask mAP$_{0.5}$**	**Mask mAP$_{0.50-0.95}$**
BoxLevelset	1333x800	COCO	73.0	65.5	44.9	64.7	43.7
Box2Mask	1333x800	COCO	79.1	78.2	61.1	79.1	57.8
SAM	N/A	SA-1B	100.0	100.0	100.0	84.3	61.4
BoxLevelset	667x400	Proprietary	86.6	77.7	57.6	79.6	48.9
Box2Mask	667x400	Proprietary	92.7	85.5	64.0	80.3	52.4

Table 3: Results after training the YOLACT-YOLOv7 model using an instance segmentation dataset. The pseudo dataset is created in a different way for each result, either pre-trained on COCO or fine-tuned on the proprietary dataset.

Pseudo labeling model	**Training dataset**	**Box mAP$_{0.50}$**	**Box mAP$_{0.50-0.95}$**	**Mask mAP$_{0.50}$**	**Mask mAP$_{0.50-0.95}$**
BoxLevelset-ResNet50	COCO	92.6	77.4	85.5	58.0
Box2Mask-ResNet101	COCO	92.5	77.4	85.6	60.3
SAM	SA-1B	94.6	78.3	87.6	65.4
BoxLevelset-ResNet50	Proprietary	92.6	76.4	85.5	52.9
Box2Mask-ResNet101	Proprietary	92.8	76.5	84.7	54.2

ing parameters of Boxlevelset are empirically fine-tuned and changed to 3.0, 3.0 and 4.0 for the focal loss, box-projection loss, and the level-set loss, respectively. For Box2Mask, these weighting parameters are changed to 4.0, 2.5, and 6.0 for the cross-entropy loss, the box-projection loss, and the level-set loss, respectively. All other training parameters remain the same for fine-tuning the models.

The results are depicted in the last two rows in Table 2. It can be seen that the detection and segmentation performances have increased with respect to the pre-trained models. Moreover, the ground-truth matching results have improved by a large margin of at least 12%, indicating that more segmentation instances are found and matched. It can be concluded that fine-tuning helps to improve the quality of the pseudo ground-truth dataset.

Figure 6 shows that more segmentation instances are found, especially for objects that are at the back of a scene. However, it also shows that the segmentation instances are often incorrect when objects are partially occluded, as shown in Figure 7. In this figure, the mask of the occluded object also includes the occluding object area.

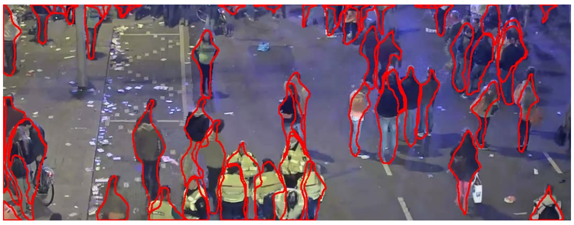


Figure 6: Generated instance labels from the fine-tuned Box2Mask model. Fine-tuning the model on the proprietary dataset resulted in greatly increased amount of labels. Blurred for privacy reasons.

4.4 Training with Pseudo Ground-Truth

In this experiment, the pseudo ground-truth datasets created in the previous experiments are used to fine-tune the YOLACT-YOLOv7 model. The fine-tuned model is then applied on the proprietary validation set. This is the last step in the method depicted in Figure 2.

The results are shown in Table 3. The highest performance is obtained when training on the dataset created using SAM. Furthermore, there is a major difference (12.5%) in the segmentation performance as seen in the last column of Table 3. This indicates that SAM has higher quality segmented instances than the box-supervised models, resulting in a better performing YOLACT-YOLOv7 model. Besides that, the YOLACT-YOLOv7 model trained on the segmentation labels generated by Box2Mask re-

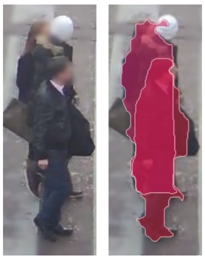
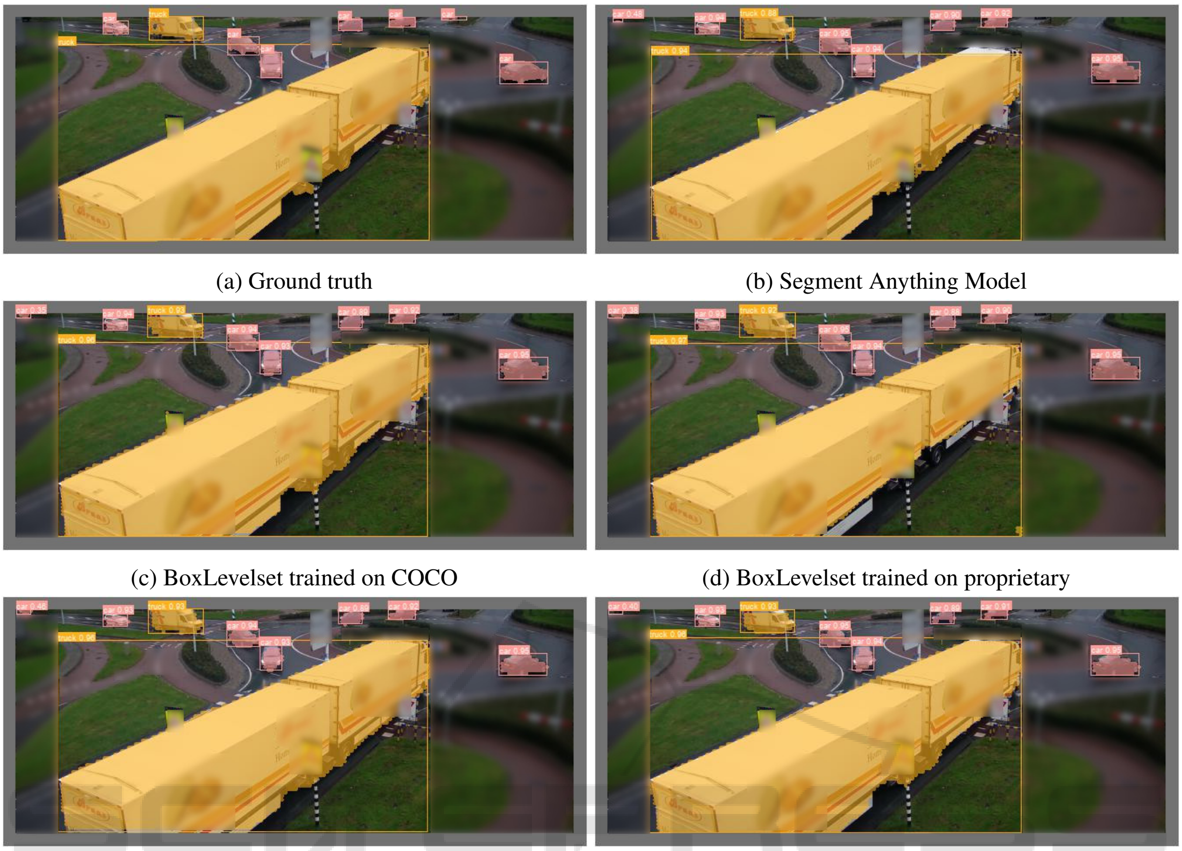


Figure 7: The left image is a snippit from the proprietary dataset, while the right shows the predicted segmentations by BoxLevelset after fine-tuning. One occluded person is not detected, and overlapping segmentations are present that are caused by the bounding box annotations.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS



(a) Ground truth

(b) Segment Anything Model

(c) BoxLevelset trained on COCO

(d) BoxLevelset trained on proprietary

(e) Box2Mask trained on COCO

(f) Box2Mask trained on proprietary

Figure 8: Results after training the YOLACT-YOLOv7 model with different generated datasets. Visually, a similar performance can be seen between all models. The main differences can be seen in boundary details. Blurred for privacy reasons.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

sults in a higher performance than the model trained on the BoxLevelset-based dataset. Surprisingly, using the datasets generated by the box-supervised models do not result in a better performing YOLACT-YOLOv7 model (bottom two rows in Table 3). This is unexpected, since the previous experiment has shown that the fine-tuned models obtained better detection and segmentation results than the models that were pre-trained on COCO. This is probably caused by errors in the segmentation masks of occluding objects, as already visually observed in the previous experiment and shown in Figure 7. Furthermore, the decrease in segmentation quality due to the reduction in input image resolution may cause inaccurate segmentation masks near object edges in the generated dataset.

For all models, the bounding-box performance is high due to model fine-tuning on the proprietary dataset. This inevitably results in the prediction of instance segmentations, since the YOLACT-YOLOv7 model simultaneously creates box and segmenation predictions, together with a joint confidence score. Hence, even though the instance segmentation masks may not be learned accurately due to false or missing ground-truth, there will always be a segmentation prediction (and a bounding-box prediction) if the confidence score is above a certain threshold.

Visual inspection shows that all YOLACT-YOLOv7 models are able to detect and segment the objects very well, even most of the participants within crowded scenes. An example image with results is shown in Figure 8, only small differences occur in these images. The object confidence is very similar for all objects. The pre-trained box-supervised results overestimate a few objects, where the fine-tuned box-supervised models underestimate the objects edges. SAM and BoxLevelset also have problems to segment the whole large object in front.

In conclusion, the box mAP scores are sufficiently high for all trained YOLACT-YOLOv7 models (over 92%). The highest mask mAP score is obtained by the YOLACT-YOLOv7 model trained on the segmentation dataset that is generated by SAM (over 87%).

5 CONCLUSIONS

This paper investigates object detection and instance segmentation for real-time traffic surveillance applications. To this end, we have adopted existing instance segmentation models by training them on a proprietary dataset. Since instance-segmentation annotations are not available for this dataset, two novel methods are proposed for generating these annotations in a semi-automated procedure. The first procedure utilizes existing pre-trained models, while the second procedure employs box-supervised models that are first finetuned on the proprietary dataset.

The YOLACT-YOLOv7 model is evaluated as optimal for traffic surveillance applications because of its high performance and low latency. Fraction training experiments on the COCO dataset show that 90% of the instance segmentation performance can be achieved when only 70% of the dataset contains instance segmentation annotations. Besides this, the YOLACT-YOLOv7 detection and segmentation performance significantly increases when it is trained on the proprietary dataset containing automatically generated instance segmentations. The instance-segmentation performance is highest when YOLACT-YOLOv7 is trained on the segmentation dataset that is generated by the Segment Anything Model (87.6% mAP). Finetuning of a box-supervised model to generate the instance segmentation ground-truth for the proprietary dataset does not result in a higher performance (85.5% mAP for BoxLevelSet). Visual inspection of the results show that future research should focus on improving instance segmentation for partially occluded objects, for example by improving the quality of the automatically generated dataset even more.

Training YOLACT-YOLOv7 on a segmentation dataset that is annotated semi-automatically forms an attractive solution, since it requires low manual annotation effort while the quality of the generated data is suitable for training. The trained YOLACT-YOLOv7 model achieves high detection and instance segmentation performance of 94.6% and 87.6% respectively, while maintaining real-time inference speed.

REFERENCES

Bolya, D., Zhou, C., Xiao, F., and Lee, Y. J. (2019). Yolact: Real-time instance segmentation. *2019 IEEE/CVF ICCV*, pages 9156–9165.

Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., and Yan, Y. (2020). Blendmask: Top-down meets bottom-up for instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8570–8578.

Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., and Schwing, A. G. (2021). Mask2former for video instance segmentation. *CoRR*, abs/2112.10764.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Getreuer, P. (2012). Chan-vese segmentation. *Image Processing On Line*, 2:214–224.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *2017 IEEE Int. Conf. on Comp. Vision (ICCV)*, pages 2980–2988.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. (2023). Segment anything.

Kuhn, H. (2012). The hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2.

Li, W., Liu, W., Zhu, J., Cui, M., Hua, X.-S., and Zhang, L. (2022a). Box-supervised instance segmentation with level set evolution. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., editors, *Computer Vision – ECCV 2022*, pages 1–18, Cham. Springer Nature Switzerland.

Li, W., Liu, W., Zhu, J., Cui, M., Yu, R., Hua, X., and Zhang, L. (2022b). Box2mask: Box-supervised instance segmentation via level set evolution. *arXiv*.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Doll'a r, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Munawar, M. R. and Hussain, M. Z. (2023). Train yolov7 segmentation on custom data.

Sharma, R., Saqib, M., Lin, C. T., and Blumenstein, M. (2022). A survey on object instance segmentation. *SN Computer Science*, 3(6):499.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475.

Wang, X., Kong, T., Shen, C., Jiang, Y., and Li, L. (2020a). SOLO: Segmenting objects by locations. In *Proc. Eur. Conf. Comp. Vision (ECCV)*.

Wang, X., Zhang, R., Kong, T., Li, L., and Shen, C. (2020b). Solov2: Dynamic and fast instance segmentation. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*.

Zwemer., M. H., Scholte., D., Wijnhoven., R. G. J., and de With., P. H. N. (2022). 3d detection of vehicles from 2d images in traffic surveillance. In *Proceedings of the 17th International Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and App. - Volume 5: VISAPP,*, pages 97–106. INSTICC, SciTePress.