Real-Time Instance Segmentation of Pedestrians using Transfer Learning*

Bare Luka Žagar, Tobias Preintner, Alois C. Knoll
Chair of Robotics, Artificial Intelligence and Real-Time Systems
Technical University of Munich
Munich, Germany
{bare.luka.zagar@tum.de, tobias.preintner@tum.de, knoll@in.tum.de}

Ekim Yurtsever
Center for Automotive Research
The Ohio State University,
Columbus, OH 43212, USA
yurtsever.2@osu.edu

***Abstract*—Real-time instance segmentation of pedestrians presents a critical core task within an automated driving pipeline. Recent research focuses on existing real-world datasets to train their instance segmentation networks. However, due to the limited size of real-world datasets, they tend to either overfit or lack accuracy. Therefore, these networks remain useless for real-world applications. Hence, we introduce a transfer learning strategy by combining a large-scale synthetic dataset and a real-world dataset for instance segmentation of pedestrians. We showcase our approach on three state-of-the-art real-time instance segmentation methods: (1) YOLACT++, (2) SipMask, and (3) BlendMask. Finally, we provide a quantitative and qualitative evaluation of our introduced approach on two publicly available urban street scenes datasets, i.e. the real-world Cityscapes dataset and the synthetic Synscapes dataset.**

***Index Terms*—Real-Time Instance Segmentation, Transfer Learning, Automated Driving Systems, Synthetic Data**

Fig. 1: Comparison of a synthetic and a real-world image. **Left**: a real-world image of an urban street scene from the Cityscapes [14] dataset. **Right**: a synthetic image of an urban street scene from the Synscape [15] dataset.

I. Introduction

Automated driving [1] could bring improved roadway safety and thus fewer vehicle accidents, according to a report [2] by the National Science Technology Council and the U.S. Department of Transportation. Moreover, the development of automated vehicle technologies (AVT) becomes more crucial if we consider that more than 94% of road accidents are due to human error [3]. Therefore, one essential safety aspect of AVT is the capability of accurate and real-time segmentation of pedestrians.

Instance segmentation [4], [5] is a core problem in the computer vision research field, and it can be seen as a combination of object detection and semantic segmentation [5]. Hence, instance segmentation mitigates the limitations of object detection and semantic segmentation, as it will not suffer from overlapping bound boxes or merged segmentation masks. Most of the recent research in instance segmentation [6]–[10] does not consider the aspect of real-time capabilities, which is crucial for any AVT system. There are only a handful of instance segmentation methods specifically designed for real-time usage [11]–[13]. However, there are no publicly available experimental results regarding pedestrian segmentation in street scenes, such as on the well-known Cityscapes urban street dataset [14], using the before-mentioned real-time instance segmentation methods.

Real-world datasets are usually relatively small in size due to the difficult and time-consuming manual annotation procedure. Hence, the training of real-time instance segmentation models on a real-world dataset might cause overfitting or could yield mediocre if not bad results. On the other hand, synthetic datasets are larger because the generation of synthetic data is usually cheap and quick, since no manual annotation is involved. For example, the real-world urban street scene Cityscapes dataset has only 5000 annotated images, while the synthetic urban street scene Synscapes [15] dataset contains 25000 annotated images. Furthermore, state-of-the-art GPU accelareted image rendering software suites [16], [17] are able to generate photorealistic images of high quality. Such that it is rather difficult to distinguish synthetic images from real-world images, as shown in figure 1. Additionally, the advantage of these highly photorealistic images is that they have a smaller gap in the difference of the data distribution w.r.t. real-world data.

Therefore, we propose a transfer learning-based training approach to improve the overall accuracy of real-time instance segmentation models for pedestrians by leveraging the large amount of data available in the synthetic dataset [15] and the correct data distribution of the real-world dataset [14]. First, we train the state-of-the-art real-time instance segmentation

*This result is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870133.

models on the large-scale synthetic Synscape dataset. The relatively large amount of data prevents the real-time instance segmentation models from overfitting, which might be the case for using only small-scale real-world datasets. Secondly, we use the weights trained on synthetic data and fine-tune the real-time instance segmentation networks on the real-world Cityscapes dataset using different amounts of training data. Here, we wanted to see how many real-world data samples are required to outperform the real-time instance segmentation models trained solely on the entire real-world dataset.

The main contribution of our work can be summarized as the following:

- We proposed a novel training methodology based on transfer learning for real-time instance segmentation methods for pedestrians,
- A comparison of the proposed transfer learning-based training approach by using different amounts of real-world data for fine-tuning,
- Extensive experiments of three state-of-the-art real-time instance segmentation methods on the Synscape synthetic dataset and on the Cityscapes real-world dataset.

II. Related Work

A. Instance Segmentation

One of the first instance segmentation method that achieved greater success and served as a base for many other works is the Mask R-CNN [6]. The Mask R-CNN network extended its predecessor, the Faster R-CNN model [18], by adding a branch to perform pixel-wise segmentation in the predicted bounding boxes. Therefore, most of the state-of-the-art instance segmentation models are two stage methods which follow the Mask R-CNN [6] strategy. They first detect the objects and then use the bounding boxes to predict the segmentation mask [9], [10]. Hence, these methods tend to have a higher runtime, which makes them useless for safety-critical real-time applications such as pedestrian segmentation for AVT systems.

B. Real-Time Instance Segmentation

YOLACT [19] and YOLACT++ [11] is the first instance segmentation method to achieve real-time. This was achieved by performing in parallel the generation of a set of prototype masks and predicting per-instance mask coefficients. In YOLACT++ the authors increased the performance over their first method, YOLACT, by leveraging deformable convolutions in the backbone to get better feature sampling and to achieve higher robustness against scale and rotation variations. BlendMask [12], on the other hand, leverages the combination of the top-down and the bottom-up approaches. The prediction of the final segmentation masks is achieved by combining the set of base masks together with the attention masks. Finally, SipMask [13] subdivides a bounding box into subregions and introduces a spatial preservation module to retain instance-level spatial information. Compared to YOLACT, which uses only one set of coefficients for the prototype masks, SipMask requires a set of coefficients for every subdivided region within a bounding box.

C. Pedestrian Instance Segmentation

As mentioned before, most of the instance segmentation methods are based on the Mask R-CNN strategy. The same is valid for pedestrian instance segmentation approaches. The authors of [20] extend the original Mask R-CNN by including prior knowledge about the body parts' proportion of pedestrians. This enables their Part Mask R-CNN [20] model to learn more information about pedestrian instances. The method proposed in [21] focuses on occluded pedestrian instances and enhances the segmentation result by leveraging pedestrian count and proposal similarity information. However, these methods have the same limitation of not being real-time capable.

III. Methodology

Our transfer learning approach for real-time pedestrian instance segmentation leverages two important properties from the synthetic Synscapes and real-world Cityscapes datasets: 1) the large amount of data in Synscape in order to prevent overfitting, and 2) the real-world data distribution of Cityscapes to achieve better performance in a real-world scenario. Hence, our training methodology can be briefly summarized in two steps:

1) Pretraining of an instance segmentation method on the photorealistic synthetic Synscape dataset.
2) Fine-tuning on the Cityscapes dataset using different amounts of training data samples.

A. Training Approach

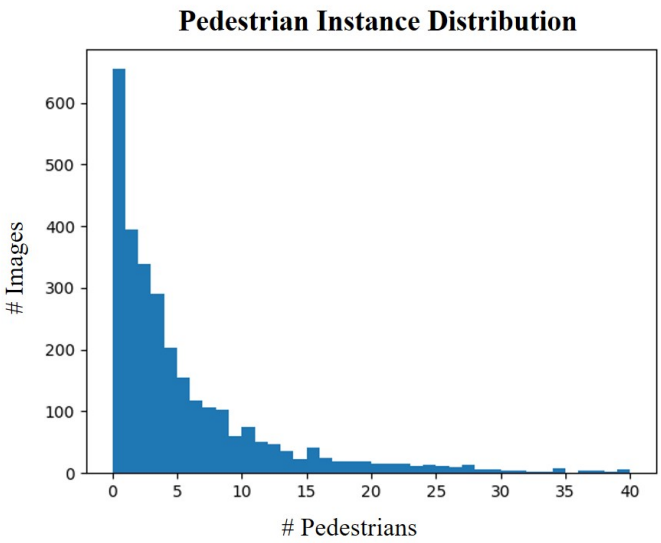

Fig. 2: The amount of images per number of pedestrian instances in the cityscapes dataset.

The introduced transfer learning-based training pipeline for real-time instance segmentation of pedestrians is shown in figure 2. First, we pre-train the real-time instance segmentation methods, i.e. YOLACT++ [11], SipMask [13], and BlendMask

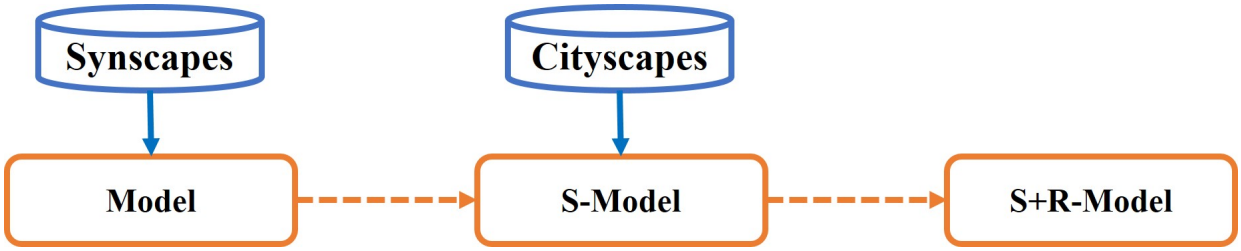

Fig. 3: Our presented transfer learning approach. First, the baseline instance segmentation method is trained using the photorealistic synthetic Synscape [15] dataset. We add to the model name the prefix "S" to indicate that it was trained on a synthetic dataset. Then, we finetune the instance segmentation model trained on synthetic data with the real-world Cityscapes [14] dataset. We add to the final model name the prefix "S+R" to indicate that it was pretrained on synthetic data and finetuned on real-world data.

[12], on the Synscape synthetic dataset using only the pedestrian class. The next step in our training pipeline consists of using the pre-trained model weights from the training on the synthetic dataset and fine-tuning them using the real-world dataset. Moreover, we investigated the dependency of the performance of the fine-tuned models w.r.t. the number of real-world data samples were used for training. Therefore, we calculated the Cityscapes training set statistics regarding the pedestrian instance distribution, as depicted in figure 2. The pedestrian instance distribution gives the information about the number of images w.r.t. the number of pedestrian instances. For example, it is interesting to observe that there are more than 600 images which contain zero pedestrian instances. Moreover, the amount of images with less than three pedestrian instances is relatively high. Therefore, we decided to filter out all the images from the training dataset, which contained less than three pedestrian instances. This filtering step equalizes the pedestrian instance distribution better throughout the training set. After applying this filtering step, the training set has approximately 1500 images left. Hence, we chose to fine-tune the real-time instance segmentation models, pre-trained on the Synscapes dataset, using 50, 500, 1000, and 1500 randomly selected training data samples.

B. Implementation Details

We evaluated transfer learning-based training pipeline for real-time instance segmentation of pedestrians using the original repositories of YOLACT++ [11], SipMask [13], and BlendMask [12]. Moreover, we followed the author's recommendation regarding the hyperparameter settings. We trained YOLACT++ with SGD using a learning rate of 0.001, a weight decay of 0.0005, and a momentum of 0.9. SipMask and BlendMask were trained with SGD using a learning rate of 0.01, while the other hyperparameters were the same as for YOLACT++. The batch sizes for YOLACT++ and BlendMask were set to 16, while for SipMask only a batch size of 8 fitted to the memory. All three methods employed the ImageNet pre-trained ResNet-50 [22] backbone. We trained all the methods for about 75000 iterations on our Workstation PC with an NVIDIA Quadro GV100 GPU and Ubuntu 20.04 LTS.

C. Evaluation Metrics

The metric used for the evaluation of the different baselines and fine-tuned models is the well-known average precision (AP), which was introduced by [23]. Generally speaking, the AP can be defined as finding the area under the precision-recall curve:

$$AP = \int_0^1 p(r)dr \tag{1}$$

where p and r denote the precision and recall, respectively. The precision and recall are defined as:

$$p = \frac{TP}{TP + FP} \tag{2}$$

$$r = \frac{TP}{TP + FN} \tag{3}$$

where TP denotes the true positives, FP the false positives, and FN the false negatives. An instance segmentation mask prediction is considered to be TP if the intersection over union (IoU) is greater than a given threshold. The IoU can be simply defined as the calculation of the overlap between the predicted mask and its ground truth pair:

$$IoU = \frac{\mid R_p \cap R_g \mid}{\mid R_p \cup R_g \mid} \tag{4}$$

where R_p and R_g denote the predicted and ground truth mask region, respectively. We follow previous works [11]–[13], where a common way of evaluating instance segmentation methods is to use the overall AP, the AP at $IoU > 0.5$ (AP_{50}), and the AP at $IoU > 0.75$ (AP_{75}).

IV. Evaluation

In this section, we begin with a detailed description of the used datasets. Then, we provide a quantitative and qualitative comparison of the real-time instance segmentation baselines w.r.t. the models trained using our proposed transfer learning-based training strategy using different amount of training data samples during fine-tuning.

A. Datasets

All experiments were conducted using the real-world Cityscapes [14] dataset and the synthetic Synscapes [15] dataset. Both datasets contain images of urban street scenes and are among the most relevant datasets to be used for AVT development. The Cityscapes dataset was created in coorperation with the Daimler AG RD Department. The images were captured in 50 cities located mostly in Germany. The Cityscapes dataset contains 5000 fine and 20000 coarse annotated images. We used the 5000 fine annotated images in our experiments since only the fine annotations are suitable for the instance segmentation task. The Synscape dataset contains 25000 photorealistic images. The dataset did not generate the images by using a driven path through the virtual environment. Instead, the authors from Synscape [15] created a unique scene for each image, which tremendously increases the variety of the data distribution. For our experiments with the Synscape dataset, we used a split of 24000 data samples for training and 1000 for validation.

B. Quantitative Evaluation

Table I shows the comparison of the baselines trained solely on the Cityscapes training set or the synthetic Synscape dataset w.r.t. to the models trained with our training strategy. We can observe that models of YOLACT++ and BlendMask trained on the Cityscapes dataset perform better w.r.t. the models trained solely on the synthetic Synscape dataset. However, one interesting observation to make, is that the SipMask model trained on the synthetic dataset performs better by a large margin, compared to the model trained on the real-world dataset. The reason for such a large discrepancy is most probably due to the relatively larger network model. Hence, the SipMask model requires significantly more training data to increase its performance.

On the other hand, it is surprising to observe that for all of the used real-time instance segmentation models trained on synthetic data in the first training step, fine-tuning on only 50 real-world images from the Cityscapes dataset is enough to be on par or outperform the models trained solely on the entire Cityscapes dataset. Moreover, we see a significant performance boost for the SipMask model pretrained on the Synscapes dataset and fine-tuned on 50 real-world images from the Cityscapes dataset. The reason for this high increase in performance can be explained by the larger amount of training data used in the first training step of our presented transfer learning-based training strategy.

Furthermore, it is evident that the best models of our proposed training strategy approach outperform by a large margin the baselines trained solely on the relatively small real-world Cityscapes dataset. Hence, the YOLACT++ model trained using our approach increases the overall AP by $+2.6\%$ compared w.r.t. to its baseline. Further, the transfer learning-based training strategy applied on the SipMask model boosts its performance by a large margin of $+9.9\%$. Finally, our training strategy improves the BlendMask model by $+4.1\%$ w.r.t. to its baseline trained only on real-world data.

TABLE I: Evaluation results on the Cityscapes validation set (pedestrian class only). The performance is reported by using the overall AP, AP_{50}, and the AP_{75} metric.

Method	AP	Δ_{AP}	AP_{50}	$\Delta_{AP_{50}}$	AP_{75}	$\Delta_{AP_{75}}$
R-YOLACT++	10.2		26.3		5.5	
S-YOLACT++	7.9		19.5		4.4	
S+R_{50^*}-YOLACT++	9.9	-0.3	27.3	+1.0	4.9	-0.6
S+R_{500^*}-YOLACT++	11.9	+1.4	30.8	+4.5	6.0	+0.5
S+R_{1000^*}-YOLACT++	12.3	+2.1	29.9	+3.6	7.8	+2.3
S+R_{1500^*}-YOLACT++	**12.8**	+2.6	**30.9**	+4.6	**8.3**	+2.8
R-SipMask	1.6		7.0		0.1	
S-SipMask	9.8		28.9		7.2	
S+R_{50^*}-SipMask	10.0	+8.4	25.7	+18.7	5.6	+5.5
S+R_{500^*}-SipMask	10.3	+8.7	26.3	+19.3	6.3	+6.2
S+R_{1000^*}-SipMask	11.1	+9.5	28.1	+21.1	6.6	+6.5
S+R_{1500^*}-SipMask	**11.5**	+9.9	**28.9**	+21.9	**7.2**	+7.1
R-BlendMask	14.3		33.6		10.4	
S-BlendMask	14.0		32.5		9.2	
S+R_{50^*}-BlendMask	14.8	+0.5	33.6	+0.0	11.6	+1.2
S+R_{500^*}-BlendMask	17.8	+3.5	39.1	+5.5	14.2	+3.8
S+R_{1000^*}-BlendMask	**18.5**	+4.2	**40.0**	+6.4	**15.1**	+4.7
S+R_{1500^*}-BlendMask	18.4	+4.1	39.4	+5.8	14.9	+4.5

* R_X denotes the training iteration with X number of training samples used from the Cityscapes training set.

The reason for the general increase on all three state-of-the-art real-time instance segmentation models, i.e. YOLACT++, SipMask, and BlendMask, lies in the proposed training pipeline. The first step in our training pipeline enables the training of the real-time instance segmentation models on a large scale dataset, which prevents the models from overfitting. Moreover, due to the photorealism of the images in the Synscape dataset, we can safely assume that the data distribution difference w.r.t. the real-world Cityscapes dataset is relatively small. Hence, the weights trained in the first step present a very good initial state for the second training stage on real-world data.

TABLE II: Inference time of the used real-time instance segmentation models.

Method	Backbone	Input size	Time (ms)
YOLACT++*	ResNet-50	640×480	34.5
SipMask*	ResNet-50	640×480	33.8
BlendMask*	ResNet-50	640×480	36.7

* The inference time is valid for all the different trained models.

Finally, we report the inference time of YOLACT++, SipMask, and BlendMask in table II. It is clear that all the methods are capable of real-time inference under the defined settings.

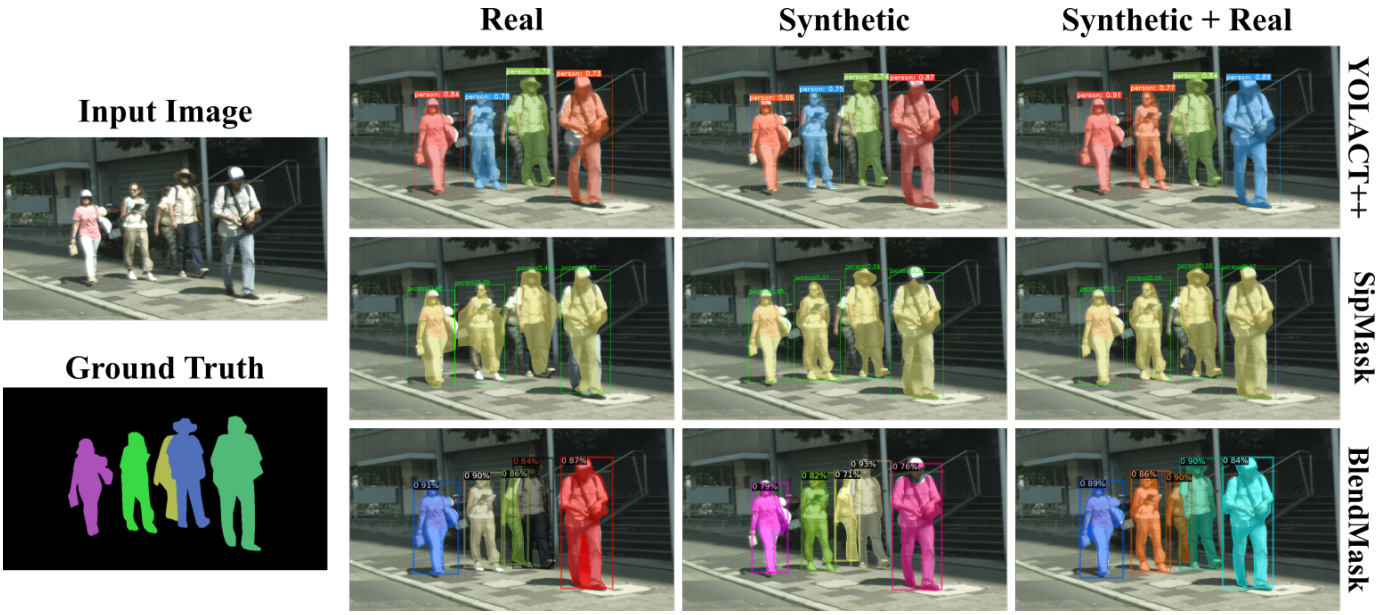

Fig. 4: The comparison of our proposed transfer learning training strategy w.r.t. to training on solely real-world or synthetic data. Three state-of-the-art real-time instance segmentation methods were used: YOLACT++, SipMask, and BlendMask. Only the best results of the models obtained from our training strategy are reported.

C. Qualitative Evaluation

The qualitative comparison of the real-time instance segmentation models trained with our training strategy and their baselines is shown in figure 4. Again, it is clearly visible that our presented transfer learning-based training pipeline visually achieves the best results. Moreover, if we take a detailed look at the qualitative results, we can observe that the instance mask predictions are more detailed for the models which were obtained with our training pipeline. For example, the models trained solely on real-world data tend to include the space between the legs of a person in the mask prediction. Contrary, the methods trained with our two-stage training approach manage to correctly predict the space between the legs of a person as background. We can assign this to the synthetic data pre-training step, because synthetic datasets have perfect annotations. Unlike the perfect annotations in synthetic data, real-world datasets suffer from faulty ground truth segmentation masks due to human error during the annotation process. Finally, we can state that our transfer learning-based training strategy mitigates the drawbacks of solely using relatively small-scale real-world datasets and increases the qualitative performance of the used real-time instance segmentation methods by a large margin.

V. Discussion and Future Work

The introduced transfer learning training strategy combining synthetic and real-world data helps to boost the performance of real-time instance segmentation methods on pedestrians. Furthermore, it is interesting to observe that even a small amount of real-world data is sufficient to fine-tune the models trained on synthetic data to achieve on par or even better results. We believe that the presented approach could help for less investigated use cases, such as manufacturing or assembly processes, where mostly no public datasets are available. To emphasize our premise, we applied the best performing model, i.e. BlendMask, to a robotic workcell use case for human intrusion detection, as shown in figure 5. Even though the model never saw such data during training, it is able to segment the human entering the robot workspace remarkably well.

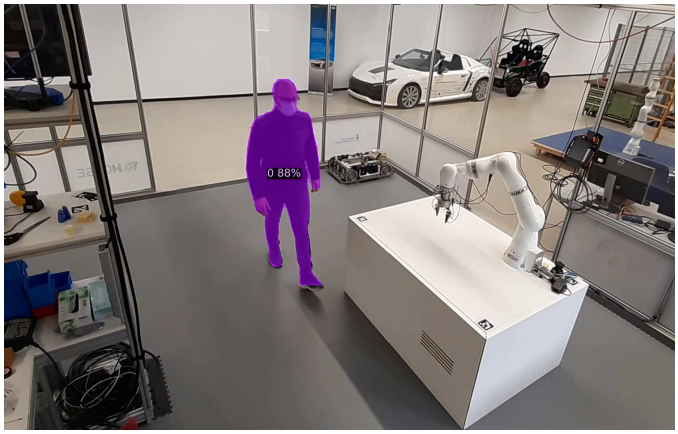

Fig. 5: Qualitative evaluation of S+R_{1000*}-BlendMask on a scene containing a robotic workcell with a Kuka LBR Iiwa and a human entering the robot's workspace.

However, the presented approach still requires time-consuming manual annotation of real-world data. To mitigate this issue, we plan to investigate unsupervised domain adaptation (UDA) methods, such as [24], which brings the learned source data distribution closer to the unlabeled target data distribution.

VI. Conclusion

In this work, we introduced a simple yet effective transfer learning-based training approach for real-time instance segmentation of pedestrians. The core concept of the introduced approach was to leverage the advantages of large-scale synthetic and real-world datasets. Therefore, we investigated the use of a two-stage training pipeline, where the models were trained using the synthetic Synscape dataset in the first step and then fine-tuned using the real-world Cityscapes dataset. The first training step provided good weight distribution due to a large amount of available training data samples in the synthetic dataset. Additionally, this prevented the models from overfitting. These pre-trained weights were then used in the second training step on real-world data. Moreover, the conducted experiments showed that only a small amount of real-world data was necessary for fine-tuning the models and outperforming their baselines.

Our work showcases the advantages of a combined training approach of synthetic and real-world datasets:

- With the proposed training approach, the amount of required real-world data can be reduced, and thus the effort of creating such a dataset based on manual annotations.
- Big neural network architectures that require a large amount of training data can be successfully trained without overfitting by using our approach.
- The introduced approach visually improves the mask prediction by leveraging the fine-grained mask annotations from synthetic data.

Finally, our transfer learning-based training strategy can help in the development of safety-critical features in automated driving systems, as it showcases a large performance boost on real-time instance segmentation of pedestrians.

References

[1] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, 2020.

[2] "Ensuring american leadership in automated vehicle technologies: Automated vehicles 4.0 (av 4.0)," National Science Technology Council and U.S. Department of Transportation, 2020.

[3] S. Singh, "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," U.S. Department of Transportation, 2018.

[4] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: state of the art," *International journal of multimedia information retrieval*, vol. 9, no. 3, pp. 171–189, 2020.

[5] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[7] H.-S. Fang, J. Sun, R. Wang, M. Gou, Y.-L. Li, and C. Lu, "Instaboost: Boosting instance segmentation via probability map guided copy-pasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 682–691.

[8] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," in *European Conference on Computer Vision*. Springer, 2020, pp. 649–665.

[9] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 213–10 224.

[10] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

[11] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact++: Better real-time instance segmentation," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[12] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "Blendmask: Top-down meets bottom-up for instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8573–8581.

[13] J. Cao, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Sipmask: Spatial information preservation for fast image and video instance segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–18.

[14] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[15] M. Wrenninge and J. Unger, "Synscapes: A photorealistic synthetic dataset for street scene parsing," *arXiv preprint arXiv:1810.08705*, 2018.

[16] B. O. Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018, accessed on Jun. 12, 2022. [Online]. Available: http://www.blender.org

[17] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[19] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9157–9166.

[20] H. Chu, H. Ma, and X. Li, "Pedestrian instance segmentation with prior structure of semantic parts," *Pattern Recognition Letters*, vol. 149, pp. 9–16, 2021.

[21] J. Xie, H. Cholakkal, R. M. Anwer, F. S. Khan, Y. Pang, L. Shao, and M. Shah, "Count- and similarity-aware rcnn for pedestrian detection," in *The European Conference on Computer Vision (ECCV)*, 2020.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, pp. 98–136, 2014.

[24] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9924–9935.