Instance Segmentation and Detection of Children to Safeguard Vulnerable Traffic User by Infrastructure

Shiva Agrawal[1][a], Savankumar Bhanderi[1][b], Sumit Amanagi[1][c], Kristina Doycheva[2][d]
and Gordon Elger[1,2][e]

[1] *Institute for Innovative Mobility (IIMo), Technische Hochschule Ingolstadt, Germany*

[2] *Fraunhofer IVI, Applied Center Connected Mobility and Infrastructure, Ingolstadt, Germany*

Keywords: Child and Adult Detection, Classification, Intelligent Roadside Infrastructure, Image Segmentation, Mask-RCNN, Traffic Flow Optimization, Transfer Learning.

Abstract: Cameras mounted on intelligent roadside infrastructure units and vehicles can detect humans on the road using state-of-the-art perception algorithms, but these algorithms are presently not trained to distinguish between human and adult. However, this is a crucial requirement from a safety perspective because a child may not follow all the traffic rules, particularly while crossing the road. Moreover, a child may stop or may start playing on the road. In such situations, the separation of a child from an adult is necessary. The work in this paper targets to solve this problem by applying a transfer-learning-based neural network approach to classify child and adult separately in camera images. The described work is comprised of image data collection, data annotation, transfer learning-based model development, and evaluation. For the work, Mask-RCNN (region-based convolutional neural network) with different backbone architectures and two different baselines are investigated and the perception precision of the architectures after transfer-learning is compared. The results reveal that the best performing trained model is able to detect and classify children and adults separately in different road scenarios with segmentation mask AP (average precision) of 85% and bounding box AP of 92%.

1 INTRODUCTION

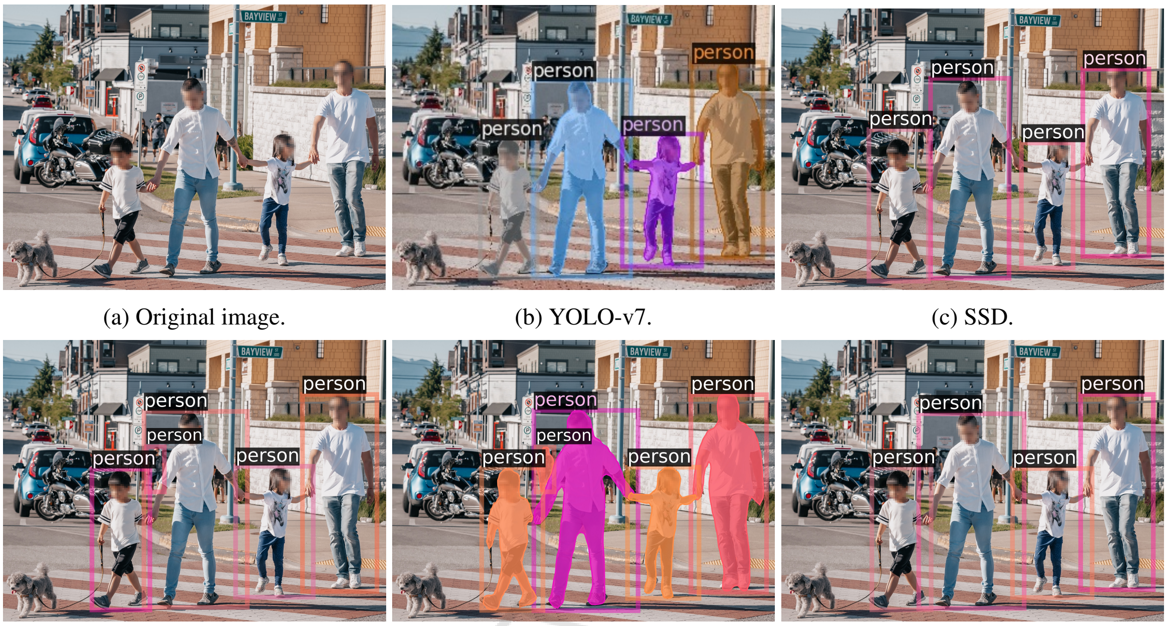
Intelligent roadside infrastructure units are usually comprised of one or more sensors to detect, classify and predict the motion of various road users. Among all the sensors, the camera is a very important sensor because it has the unique ability to detect different colours, shapes, sizes, textures, and types of objects. Hence, it is widely used to classify various road users like pedestrians, bicycles, motorbikes, cars, trucks, buses, animals, static objects, etc. This classification helps intelligent roadside infrastructure units to decide critical and non-critical situations arising on the road. For example, if a pedestrian is detected crossing the road when the traffic lights of the vehicle lane are red, then it is normal condition but if the pedestrian is detected crossing the road when the traffic lights of the vehicle lane are green, then this is a critical situation. In such a condition, an intelligent roadside infrastructure unit has to immediately send a warning signal to passing vehicles in order to avoid accidents and save human lives. Such a situation may not rise so often when an adult is crossing the road but it is often possible when a child is crossing the road.

A child may not follow or understand all the traffic rules. Also, a child may cross the road anytime or may stop or start playing in the middle of the road. Hence, it is important that intelligent roadside infrastructure units are able to detect and classify pedestrians, separately as an adult or a child to increase their safety on the road and also to avoid accidents. A camera can recognize various road users better than other sensors, so it is wise to use it to recognize a human (or pedestrian) as a child or an adult.

The results from traditional computer vision algorithms in this domain are limited but artificial in-

[a] https://orcid.org/0000-0001-8633-341X
[b] https://orcid.org/0000-0001-7257-6736
[c] https://orcid.org/0000-0003-0132-8115
[d] https://orcid.org/0000-0002-3340-7048
[e] https://orcid.org/0000-0002-7643-7327

Agrawal, S., Bhanderi, S., Amanagi, S., Doycheva, K. and Elger, G.
Instance Segmentation and Detection of Children to Safeguard Vulnerable Traffic User by Infrastructure.
DOI: 10.5220/0011825400003479
In *Proceedings of the 9th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2023)*, pages 206-214
ISBN: 978-989-758-652-1; ISSN: 2184-495X



(a) Original image. (b) YOLO-v7. (c) SSD.

(d) Faster R-CNN. (e) Mask-RCNN. (f) Retina Net.

Figure 1: Both adults and children are detected as persons in the camera image by different state-of-the-art AI-based detectors. Figure 1a is taken from open source (Productions, 2021).

telligence (AI) based computer vision algorithms are very good at detecting and classifying various road users from camera images. The latest state-of-the-art algorithms like faster-RCNN (Region-Based Convolutional Neural Network) (Ren et al., 2015), SSD (Single Shot Detector) (Liu et al., 2016), Mask-RCNN (He et al., 2017), RetinaNet (Lin et al., 2017b) and YOLOv7 (You Only Look Once version 7) (Wang et al., 2022) are widely used for road user classification in many research and commercial applications. But as highlighted in figure 1, none of these state-of-the-art object detectors is able to separately recognize whether the detected person is an instance of a child or an adult.

The work in this paper is focused to solve this problem by using the transfer learning (Zhuang et al., 2020) approach. Among the state-of-the-art models, the Mask-RCNN model is selected for the work because it has the ability to generate instance segmentation (object mask) along with bounding box and classification output. The literature survey during the work concluded that there is no public dataset available for such a problem. Hence, images from various sources containing humans (both child and adult) are collected and annotated using a semi-automatic labelling framework. For specific images, manual labelling of the data is also performed. Thereafter, the pre-trained Mask-RCNN network is modified to adapt it for a two-class object detection and segmentation task and then trained on the dataset using six different feature extraction backbone architectures including two different baselines. All the trained models are evaluated and the best performing model is selected to use for traffic flow optimization use-case for the intelligent roadside infrastructure (Agrawal et al., 2022).

This paper is outlined as follows: section 2 provides insights into transfer learning, Mask-RCNN, and detectron2 framework. Section 3 describes the approach and method of child and adult detection and instance segmentation that includes the process of data collection, data annotation, dataset generation, AI-based model development, training, and testing. Section 4 provides the results of the proposed method and then at last conclusion is given.

2 TECHNICAL BACKGROUND WITH RELATED WORK

2.1 Transfer Learning

Transfer learning (Zhuang et al., 2020) is the approach widely used in deep learning applications. Training of deep learning architectures requires a very huge amount of labelled data. For every application, the collection of such a huge amount of data is not practically possible. With the help of transfer learning, one can use the pre-trained weights and deep learning model designed for one application to another similar application with some modifications. In this approach, the amount of new data required for training is comparatively low and can provide good results with high accuracy.

Figure 2 visually describes the concept of transfer learning. The source model is the original model from either the same domain or from another domain that is trained previously using a large amount of labelled data. The knowledge in form of trained weights and architecture of this source model is then transferred and used partially or fully to train another model, known as the target model. To train this target model, a comparatively very small amount of labelled data is required to get high-performance metrics because this target model is not trained from scratch but is trained over the available knowledge of the source model.

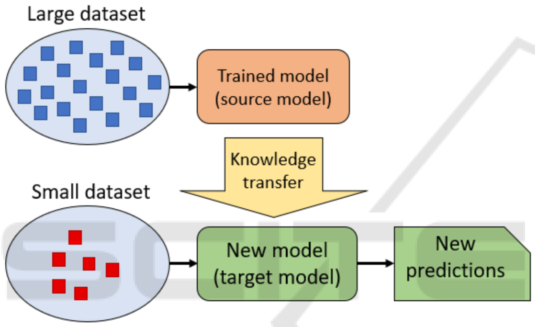


Figure 2: Overview of transfer learning approach.

Transfer learning is applied successfully in many fields including medical imaging, aerospace, natural language processing, audio processing, autonomous vehicle development, etc. As a result, the available literature in this area is very vast, and, hence only some of the work from different fields is cited here. For example, in (Liang et al., 2016) aerial images are classified using transfer learning for remote sensing image understanding tasks. The work stated in (Cao et al., 2013) used transfer learning for better and more accurate detection of pedestrians from camera images. The author in (Hu and Yang, 2011) has used this approach to identify and predict various human activities. In (Alzubaidi et al., 2020), the author has used transfer learning to classify images to detect breast cancer and in (Kocmi and Bojar, 2018) for low-resource neural network-based machine translation application. Similarly, the work stated in (Gáti and Kiss, 2021) represents sound signals as images and then uses transfer learning to classify sounds through an image classifier as the source model.

2.2 Mask-RCNN

Mask-RCNN (He et al., 2017) is a state-of-the-art neural network model which is widely used in image-based object detection and image segmentation applications. This model consists of a backbone network that does the work of feature extraction (high level to low level) from an input image. These extracted features along with pre-defined anchors are fed into a region proposal network (RPN) followed by an ROI (region of interest) alignment block to get fixed-size proposals. These proposals are then passed through a series of fully connected layers (FCN) to generate object class probabilities and bounding box regression and also passed through a series of CNN layers to predict the binary mask of each detected object. The backbone of the Mask-RCNN that is responsible for feature extraction can be used from multiple state-of-the-art classifier models. Among them, the most widely used are different variants of ResNet(residual network)(He et al., 2016) with FPN (feature pyramid network)(Lin et al., 2017a).

Mask-RCNN together with transfer learning is used in many applications. For example, the work described in (Zhang et al., 2020) used transfer learning together with Mask-RCNN to detect damaged vehicles using camera images. Authors in (Doğru et al., 2020) used the same combination of Mark-RCNN with transfer learning to develop a target model with a small amount of labelled data to detect dents on the upper body of the aircraft for automatic maintenance. Similarly, authors in (Shenavarmasouleh et al., 2021) used this combination in the medical field to detect lesian in fundus images to address the diabetic retinopathic problem.

From this literature survey, it is evident that transfer learning is a very powerful method and it has significantly generated good results by using many different deep learning models (as source models) including Mask-RCNN. Hence, in this paper, this powerful and proven combination of transfer learning with Mask-RCNN is used to develop the target model to classify children and adults as two distinct classes from camera images for intelligent roadside infrastructure applications. In addition, no direct related work found in the literature that has solved this specific issue which further motivated the proposed work of this paper.

2.3 Detectron2

Detectron2 (Wu et al., 2019) is a widely used open-source modular software framework from the AI group of Facebook. It is implemented in py-

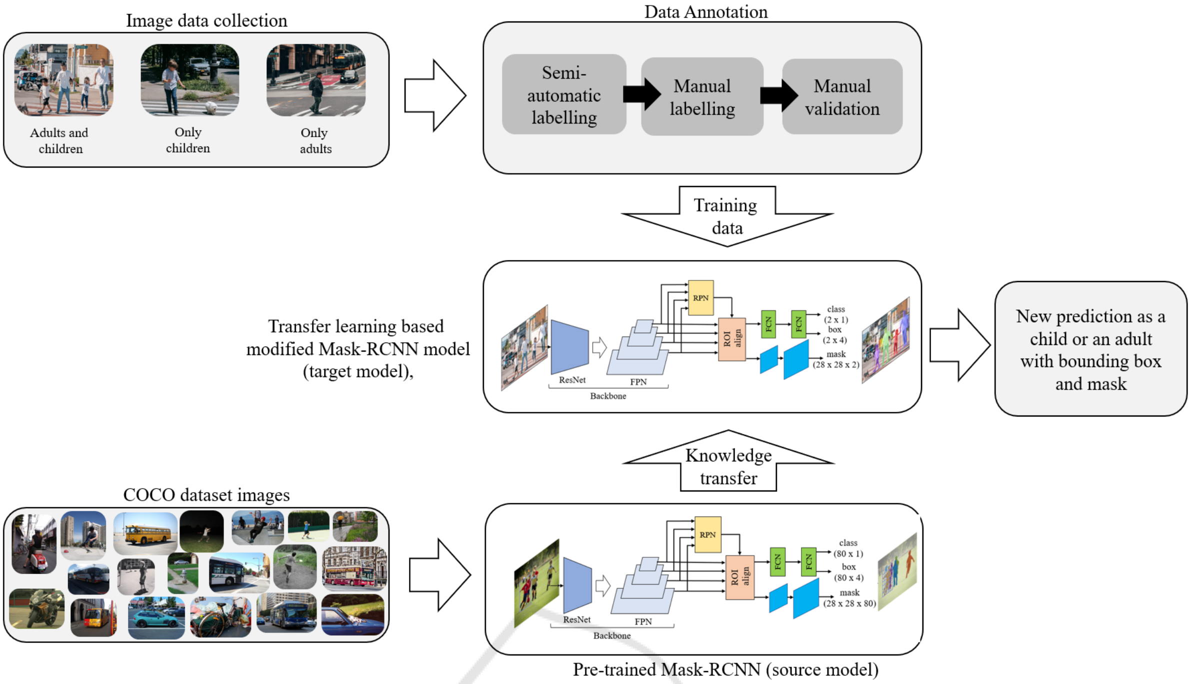


Figure 3: Proposed method for detection and instance segmentation of a child and an adult.

torch (Paszke et al., 2019), a python library for AI-based development. It provides easy to use interface and takes low computation time on single or multiple GPU systems. Detectron2 is the successor of detectron that was started with the Mask-RCNN benchmark. Hence, Mask-RCNN is available in the detectron2 framework, and with this framework, the process of using Mask-RCNN together with transfer learning is comparatively easy. Hence, in this work, Mask-RCNN within the detectron2 framework is used.

3 PROPOSED METHOD

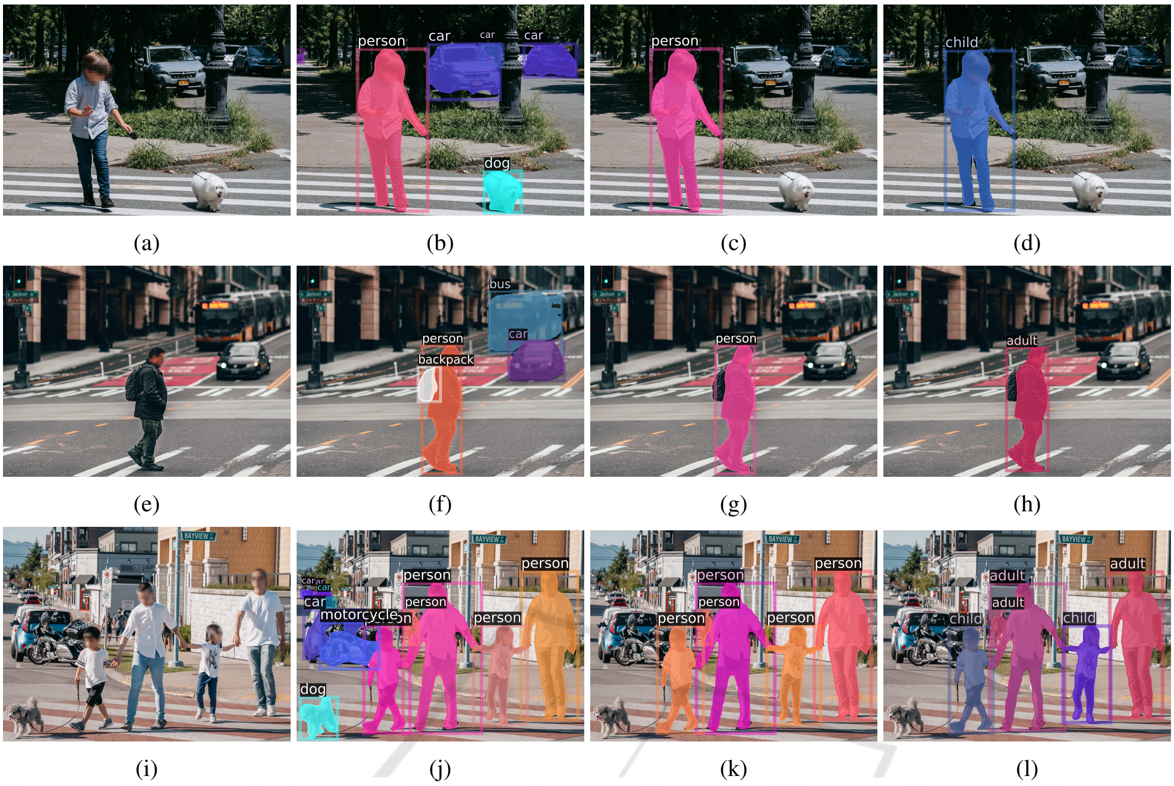
The complete pipeline of the proposed method for detection and instance segmentation of a person as a child or an adult is highlighted in figure 3. Various images of children and adults in the road environment were collected during the image data collection process. All the collected images were annotated by a semi-automatic labelling pipeline to generate a dataset for the work. The resulting dataset was then split between train and test samples for the training purpose.

In the transfer learning-based approach, the Mask-RCNN model is used as the source model that is pre-trained on COCO (common objects in context) dataset (Lin et al., 2014). The target model, i.e. the described model of the work is then selected by freezing all the initial layers of the source model and modifying the classification, bounding box regression, and mask generation layers for two categories of objects, i.e. a child and an adult. Then the target model is trained on a relatively small size dataset of child and adult images. For each feature detection backbone architecture used during the work, the target model is designed using the corresponding source model in a similar manner for training and evaluation.

3.1 Data Collection

As stated before, for this work, a comparatively small dataset is required because of the transfer learning approach, but none of the publicly available datasets has labels for a person as an adult or a child. Hence, for this work, at first, images containing instances of children and adults were collected from many open-source websites and publicly available datasets. Please note that the images from public datasets containing adults and children are available only as a person.

These images were collected in a variety of settings and poses for instances of adults and children in road environments. With further refinement, a total of 506 images were selected that were categorized as images containing instances of only children, images containing instances of only adults, and images containing instances of both children and adults.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 4: Semi-automatic labelling framework for data annotation. From (4a source - (Wolf, 2021)) to (4d) is the labelling process for images having only children. From (4e source - (Han, 2020)) to (4h) is the labelling process for images having only adults and from (4i source - (Productions, 2021)) to (4l) is the labelling process for images having both adults and children.

3.2 Data Annotation

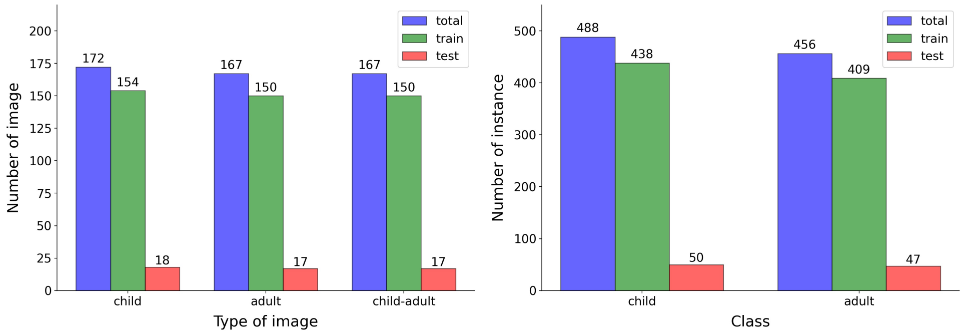
To annotate the collected images, a semi-automatic labelling framework is developed. This framework is depicted in figure 4. Please note that the figure 4d, 4h, and 4l are not the output of the final model, rather they show the label transfer from person to child or adult for generating the annotations using semi-automatic labelling framework.

At first, the state-of-the-art Mask-RCNN model that is pre-trained on the COCO dataset with 80 different classes including persons is used within the detectron2 framework to generate the bounding box, class, score, and object mask of all the 80 classes. Then only instances of a person are preserved and the rest are removed. Then depending on the type of image used as input, the label of the person is transferred as a child or an adult. For example, as shown in figure 4, images from 4a to 4d depict this complete semi-automatic labelling pipeline for images contains only child (one or more instances). The original image of figure 4a is fed into the Mask-RCNN model which results in figure 4b and then only the instances of person are kept which is shown in figure 4c. At last, the label of a person is transferred to the child as shown in figure 4d. Similarly, the pipeline for images containing only adults (one or more instances) is shown from figure 4e to figure 4h.

For the labelling of images containing instances of both children and adults, the same pipeline is used to generate labels as a person but then manually each instance in the image is checked and labels are transferred as either a child or an adult. Due to this manual intervention, the proposed framework is named a semi-automatic labelling framework. After generating the annotations through the aforementioned methodology, a manual refinement step is performed to account for the cases where the network might have failed to detect a person. This was followed by a final manual validation step and then the annotations were saved in a COCO format for training purposes.

3.3 Dataset Summary

Figure 5 illustrates the summary of the generated dataset and its distribution. The entire dataset consists of 506 images, from which 454 are selected as training samples and 52 are chosen as testing sam-



(a) Images type distribution.

(b) Class distribution.

Figure 5: Dataset sumamry.

Table 1: Configuration Parameters.

Parameter	Value	Description
base_lr	0.01	learning rate for updating the weights
batch_size	4	number of images in a batch
freeze_at	2	freeze first freeze_at layers
gamma	0.1	learning rate decay factor
steps	[5000,15000,20000]	steps to decay base_lr by gamma
optimiser	SGD with 0.9 momentum	network optimiser
max_iters	30000	maximum iterations to train
eval_period	500	evaluation every eval_peroid iterations
warmup_iters	500	number of iterations to increase the learning rate to base_lr
warmup_type	linear	ramp learning rate linearly
checkpoint	500	save weights every checkpoint iterations
train_aug	resize shortest edge, brightness saturation, contrast horizontal flip	train time data augmentations
test_aug	resize shortest edge, horizontal flip	test data augmentations

ples, by keeping a 90% - 10% train-test ratio. Further figure 5a shows the distribution of test and train samples for each type of image and figure 5b provides the instance-wise distribution of training and test data.

3.4 Development of Model

The Mask-RCNN (pre-trained on coco dataset) model (source model) is imported directly from the Detectron2 models' catalog and configuration was changed to generate detection and segmentation outputs for 2 classes to match objects of interest for the work. The mask branch of the model generates by default the class-wise mask of spatial dimensions (28,28) that is incompatible with the size of the input image. Hence, in a post-processing step, the masks are up-sampled to match the input image spatial dimensions. Apart from that, other configurations like anchor labelling, loss functions, and RPN settings are preserved from the source model. The hyper-parameters configured for the training of the models are given in Table 1.

For the target models, the first two stages of the original backbones are frozen, and only in later layers, training is carried out for child and adult detection and segmentation. The learning rate of the models is changed in steps during the training as highlighted in Table 1. At first, the learning rate (base_lr) is set to 0.01 which is reduced to 0.001 after 15000 iterations and then further reduced to 0.0001 after 20000 iterations during the training and for uniform comparison of results for all the trained models, a constant seed value of 42 is used.

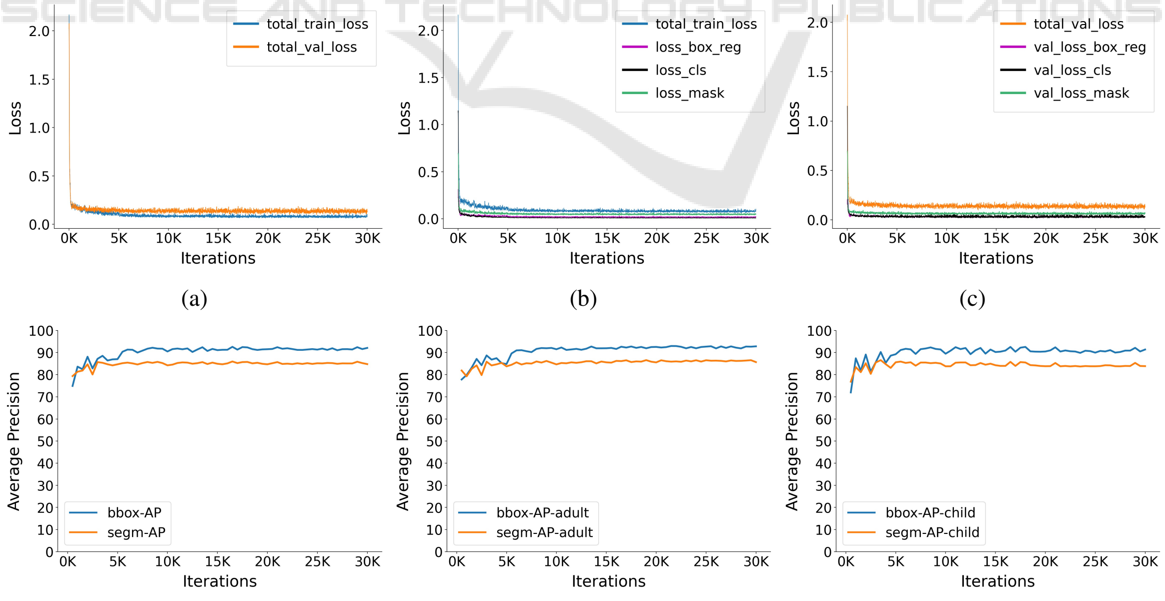
4 EXPERIMENTS AND RESULTS

Four different backbone architectures including ResNet50, ResNet101, ResNeXt101 (Xie et al., 2017), and ResNeXt152 (with cascaded Mask-RCNN head) were used as a source model during the training. Further, the ResNet50 and ResNet101 backbones are used with two different baselines as highlighted in Table 2. Backbones with baseline 1 are pre-trained with a standard 3x schedule and 120 COCO epochs, whereas the backbones with baseline 2 are pre-trained for a longer schedule of 400 COCO epochs using copy-paste augmentations as described in (Ghiasi et al., 2021). As a result, a total of six different backbones with two baselines are used in conjunction with FPN (feature pyramid network) for training and evaluation of the child and adult detection problem.

Each variant was trained independently on the machine with 16 core CPU of 3.4 GHz, 128 GB RAM, RTX 3090 GPU with 24 GB memory, and AMD Ryzon 9 5950x processor. Each variant excluding ResNeXt152 took around 5 hours for training and ResNeXt152 took around 18 hours for training. Table 2 highlights the performance in form of AP (average precision) of the bounding box and segmentation mask of all models after training with the same hyper-parameters as described in section 3.4. Results clearly show that the trained model with backbone architecture of ResNet50 + FPN with baseline 2 has the highest AP value of 92.43% for bounding box generation and 85.85% for segmentation mask. Hence, for this best-performing model, the training and validation loss, and performance metrics for each class with respect to training iterations are further given in figure 6.

Table 2: Comparison of results (each backbone is with FPN).

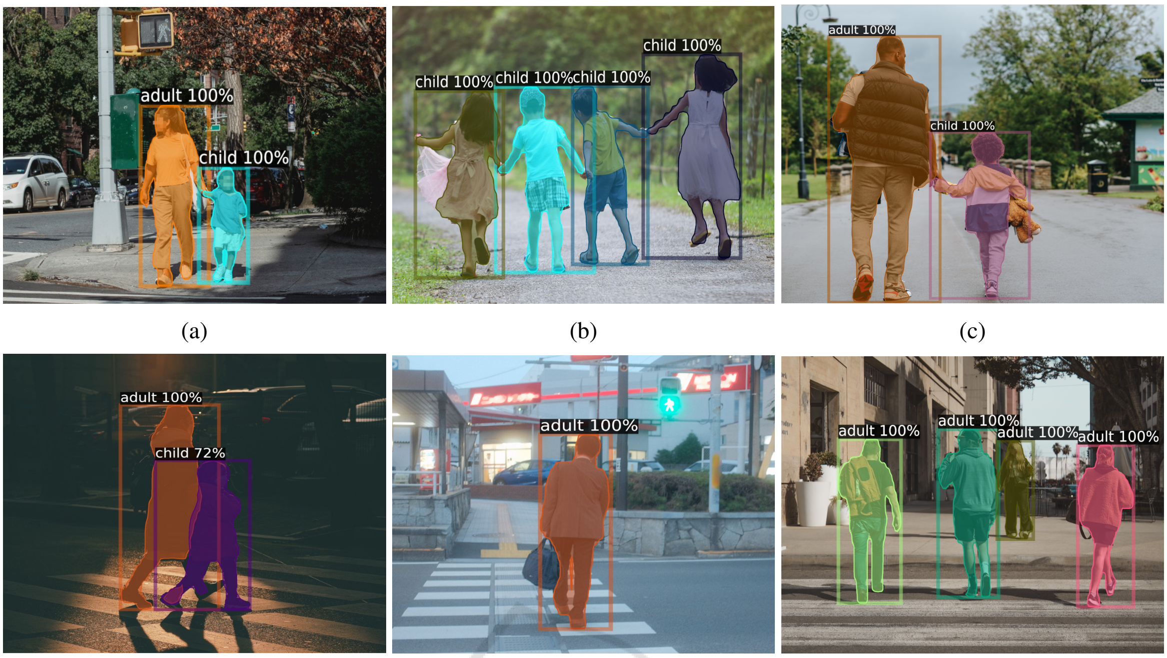
Metric	**baselines 1**				**baselines 2**	
	ResNet50	ResNet101	ResNeXt101	ResNeXt152_Cascade	**ResNet50**	ResNet101
box AP	81.54	84.05	84.10	91.57	**92.43**	90.62
mask AP	82.20	83.16	84.31	83.95	**85.85**	84.28
box AP- adult	80.65	83.41	83.26	92.07	**92.71**	91.98
box AP- child	82.53	84.70	84.92	91.06	**92.14**	89.27
mask AP-adult	81.50	81.68	82.75	83.00	**86.40**	85.37
mask AP-child	82.80	85.45	85.87	84.87	**85.30**	83.19



(a) (b) (c)

(d) (e) (f)

Figure 6: Performance and loss graphs of best performing model (Mask-RCNN with baseline 2 ResNet50). The x-axis represents iterations and the y-axis represents parameters. (6a) compares training loss and validation loss during training. Further breakdown of training loss and validation loss into sub-categories is shown in (6b) and (6c) respectively. 6d shows the average precision of the model for box and masks. per-category average precision is shown in (6e) and (6f).



(a) (b) (c)

(d) (e) (f)

Figure 7: Inference results of the work. All the original images are taken from open source websites, 7a from (Ogino, 2021), 7b from (Khac, 2020), 7c from (Lusina, 2021), 7d from (Kim, 2021), 7e from (Mak, 2021) and 7f from (Cooks, 2022).

Further analysis of the results shows that among the two models with baseline 2, ResNet50 outperforms ResNet101 in terms of all the reported performance metrics. These models are pre-trained on a longer schedule and are therefore capable of capturing very complex patterns within an image. This explains the very high AP and low loss of these models from the beginning as highlighted in figures 6d, 6e and 6f. However, the increased complexity of the ResNet101 model with baseline 2 is affecting negatively the two-category problem which results in lower AP as compared to ResNet50 with baseline 2. Additionally, when comparing variants from baseline 1, then cascade Mask-RCNN with ResNeXt152 backbone provides the highest AP for the box and the mask.

5 CONCLUSION

In the described work, an instance segmentation and detection model using an AI-based neural network is developed to separately classify and detect humans as a child or an adult. For this purpose, Mask-RCNN model with different backbone architectures of ResNet50, ResNet101, ResNeXt101, and ResNext152 cascade together with FPN including two different pre-trained baselines are used with transfer learning. From the results, it is found that the Mask-RCNN model with ResNet50 with baseline 2 (i.e. pre-trained model with 400 epochs) performs the best with segmentation mask AP (average precision) of 85% and bounding box AP of 92%.

For further work, the network will be trained by collecting boundary cases, e.g. small adults and adults in cower positions, to analyze and improve the classifier. However, also when the differentiation might not reach 100% the increase of security for children will be improved by additionally warning the ongoing traffic that children are present especially around bus stops, areas in front of schools and crossings where children passing on the school ways using intelligent roadside infrastructure.

ACKNOWLEDGEMENTS

The research work is supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy (StMWi) in the project “INFRA – Intelligent Infrastructure”.

REFERENCES

Agrawal, S., Song, R., Kohli, A., Korb, A., Andre, M., Holzinger, E., and Elger, G. (2022). Concept of smart infrastructure for connected vehicle assist and traffic flow optimization. In *VEHITS*, pages 360–367.

Alzubaidi, L., Al-Shamma, O., Fadhel, M. A., Farhan, L., Zhang, J., and Duan, Y. (2020). Optimizing the performance of breast cancer classification by employing the same domain transfer learning from hybrid deep convolutional neural network model. *Electronics*, 9(3).

Cao, X., Wang, Z., Yan, P., and Li, X. (2013). Transfer learning for pedestrian detection. *Neurocomputing*, 100:51–57. Special issue: Behaviours in video.

Cooks, J. (2022). [Online; accessed September, 2022].

Doğru, A., Bouarfa, S., Arizar, R., and Aydoğan, R. (2020). Using convolutional neural networks to automate aircraft maintenance visual inspection. *Aerospace*, 7(12):171.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., and Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2917–2927.

Gáti, N. and Kiss, A. (2021). Sound classification with transfer learning (13th joint conference on mathematics and computer science (the 13th macs), on october 1-3, 2020).

Han, R. (2020). Man walking on pedestrian lane. [Online; accessed September, 2022].

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Hu, D. H. and Yang, Q. (2011). Transfer learning for activity recognition via sensor mapping. In *Twenty-second international joint conference on artificial intelligence*.

Khac, A. (2020). A group of children walking hand in hand on unpaved road. [Online; accessed September, 2022].

Kim, R. (2021). [Online; accessed September, 2022].

Kocmi, T. and Bojar, O. (2018). Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

Liang, Y., Monteiro, S. T., and Saber, E. S. (2016). Transfer learning for high resolution aerial image classification. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–8.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2014). Microsoft coco: Common objects in context.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Lusina, A. (2021). Unrecognizable black father with son holding hands on city road. [Online; accessed September, 2022].

Mak (2021). [Online; accessed September, 2022].

Ogino, K. (2021). Asian woman and girl standing near crosswalk. [Online; accessed September, 2022].

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. pages 8024–8035.

Productions, P. (2021). Family crossing the street while holding each other's hands. [Online; accessed September, 2022].

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

Shenavarmasouleh, F., Mohammadi, F. G., Amini, M. H., Taha, T., Rasheed, K., and Arabnia, H. R. (2021). Drdrv3: Complete lesion detection in fundus images using mask r-cnn, transfer learning, and lstm. *arXiv preprint arXiv:2108.08095*.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.

Wolf, K. (2021). Kid and dog crossing the street. [Online; accessed September, 2022].

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995.

Zhang, Q., Chang, X., and Bian, S. B. (2020). Vehicle-damage-detection segmentation algorithm based on improved mask rcnn. *IEEE Access*, 8:6997–7004.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.