2021 IEEE Intelligent Vehicles Symposium (IV)
July 11-17, 2021. Nagoya, Japan

Simple Pair Pose - Pairwise Human Pose Estimation in Dense Urban Traffic Scenes

Markus Braun[1,2], Fabian B. Flohr[2], Sebastian Krebs[1,2], Ulrich Kreßel[2], and Dariu M. Gavrila[1]

***Abstract*— Despite the success of deep learning, human pose estimation remains a challenging problem in particular in dense urban traffic scenarios. Its robustness is important for follow-up tasks like trajectory prediction and gesture recognition. We are interested in human pose estimation in crowded scenes with overlapping pedestrians, in particular pairwise constellations. We propose a new top-down method that relies on pairwise detections as input and jointly estimates the two poses of such pairs in a single forward pass within a deep convolutional neural network. As availability of automotive datasets providing poses and a fair amount of crowded scenes is limited, we extend the EuroCity Persons dataset by additional images and pose annotations. With 46,975 images and poses of 279,329 persons our new *EuroCity Persons Dense Pose* dataset is the largest pose dataset recorded from a moving vehicle. In our experiments using this dataset we show improved performance for poses of pedestrian pairs in comparison with a state of the art method for human pose estimation in crowds.**

I. Introduction

A reliable detection of vulnerable road users, like pedestrians and riders, is fundamental in the context of intelligent vehicles. In particular for fully automated driving systems understanding the surroundings of the vehicle and estimating the future behavior of other traffic participants is crucial. Still, a human driver not only depends on the positions for predicting the future trajectory of vulnerable road users. Additional cues like the line of gaze, hand gestures, or the gait cycle are automatically perceived and processed. To recreate these capabilities, the pose of persons defined by the position of the joints can be used as intermediate representation for gesture recognition and intention estimation [1]. Mutual occlusions of pedestrians in crowd situations in an automotive scenario pose challenges not only for detection but also for deep learning based pose estimation [2]–[13].

This could be the case for a group of pedestrians waiting at a bus stop. As each member in such a group can suddenly step out and enter the street, a reliable detection and pose estimation is equally important for all the pedestrians in the group, even for the ones further in the back, potentially occluded by other pedestrians.

Regarding pose estimation, top-down approaches are still leading on the MSCOCO dataset [16] frequently used for benchmarking. These methods first detect all persons in an image and estimate the pose of each person in a second stage, whereas bottom-up approaches first try to find all joints within an image, which are then clustered into instances. Often, the underlying detection methods used in top-down approaches depend on a non-maximum suppression (NMS) as post processing step, to avoid multiple detections for a single instance. Interestingly, many of the top performing methods utilize a simple greedy implementation based on a single intersection over union threshold (IoU) [19]. Selecting this threshold for suppression poses a tradeoff between recall and precision [20] and leads to missing detections in particular in groups. This threshold is often set to 0.5 for pedestrians, meaning if two pedestrians have a higher mutual IoU only one will be detected assuming perfectly localized detections. Such pedestrians with a higher mutual IoU than 0.5 are defined as *pedestrian pair* throughout our work. If there are multiple pedestrians with an IoU higher than 0.5, only the two pedestrians with the highest mutual IoU are regarded as pair. The data statistics of our new EuroCity Persons Dense Pose dataset shows that pedestrian pairs are still common (5.9%), while only 0.2% of the pedestrians have a mutual IoU greater 0.5 with at least two other pedestrians. Therefore, we are mainly interested in detection and pose estimation of these pedestrian pairs. A lot of work has been published on improving the detection recall in groups [19], [21]–[24]. In our methodology, we follow the idea of [24] to jointly detect sets of pedestrians based on a single proposal. Therefore, we adapt a YOLOv3 [25] to jointly detect pairs of pedestrians.

Regarding top-down pose estimation in groups, the cropped detections used as input often contain parts of other

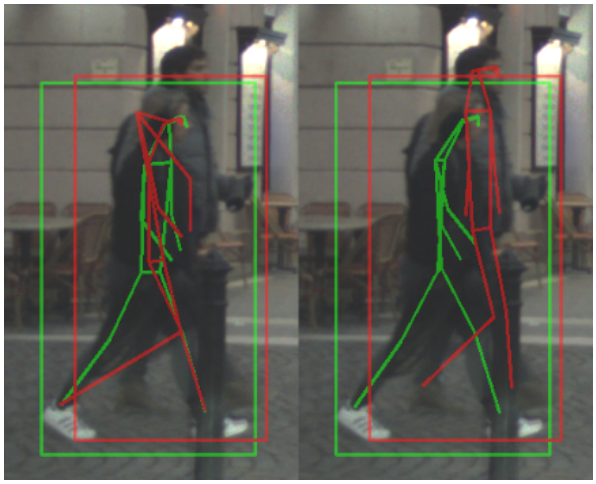

Fig. 1: Qualitative pose estimation result of AlphaPose+ [11] (left) and our Simple Pair Pose method (right) for a pedestrian pair. In our methodology and experiments we are particularly interested in such pair situations in dense urban traffic scenes.

1) Intelligent Vehicles group, TU Delft, The Netherlands
2) Environment Perception group, Mercedes-Benz AG, Germany

2021 IEEE Intelligent Vehicles Symposium (IV) | 978-1-7281-5394-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/IV48863.2021.9575435

TABLE I: Overview of human pose datasets.

Dataset	**ECPDP** (ours)	**TDUP** [14]	**PedX** [15]	**MSCOCO** [16]	**MPII** [17]	**AI Chall.** [18]	**CP** [11]	**OP** [12]
Domain	Autom.	Autom.	Autom.	General	General	General	General	General
# Images	47k	21k	5k (stereo)	200k	25k	**300k**	20k	9k
# Person Poses	279k	93k	14k	250k	40k	**700k**	80k	18k
Avg. Persons/Img	**5.9**	4.4	2.8	1.3	1.6	2.3	4.0	2.0

persons. Sometimes the target pose becomes ambiguous [13], in particular when the overlap of persons is very high as in our pair situations. [13] solves the disambiguation by adding an additional input hint for the target pose, while other methods optimize poses of multiple persons in a post-processing step [11], [12]. In our pose estimation approach, we make use of paired detections and jointly estimate the two poses within a single network. Thus, we solve the disambiguation of poses by training separate experts for front pedestrians and back pedestrians in pairs. For an exemplary result see Figure 1.

Recently, the EuroCity Persons (ECP) benchmark dataset [20] has been published to advance progress regarding bounding box based person detection in automotive scenarios. It consists of images recorded with a front facing camera mounted behind the windshield of a moving vehicle. We extend this dataset by additional images from two side facing cameras that have been synchronously recorded. For the selection of the 47k images that form our new EuroCity Persons Dense Pose (ECPDP) dataset we focus on crowded scenes. To enable the advancement in the field of pose estimation, we provide annotations for all 17 joints of the pedestrians and riders. See Table I for an overview of human pose datasets.

II. Related Work

a) Multi Person Detection: Most detection approaches can be clustered into two-stage approaches [26], [27] or one-stage approaches [25], [28]. The non-maximum suppression (NMS) used in a post processing step to suppress multiple detections per object poses a tradeoff between recall and precision [20].

There are several approaches to improve the recall in particular in crowd situations, without losing precision. In Soft-NMS [19] detections are not discarded, but their class score is reduced, if they overlap with another detection, that has a higher confidence. [21] proposes a network architecture to learn the NMS task using bounding box locations and class scores as input. Thus, the NMS could be trained in a fully end-to-end detection framework. In [22] a density value is estimated per prediction that is used instead of the single IoU threshold within the greedy NMS. A high density value leads to less suppression and a higher recall in groups. [23] builds upon this idea and additionally estimates a diversity value. This discriminative diversity value is estimated in an embedded feature space and is fed into the adapted NMS algorithm. In [29], a special loss coined Repulsion Loss is used, to push detections of separate instances away from each other to lower the IoU between such detections. [24] tries to detect all objects in a group based on a single proposal. These set detections do not suppress each other within the NMS.

b) Multi Person Pose estimation: Multi person pose estimation can be clustered into bottom-up and top-down approaches. Bottom-up approaches [2]–[6] first try to find all joints within an image, which are then clustered into instances. Early approaches solve the clustering by integer linear programming [2], [3]. In [4] part affinity heatmaps are estimated in addition to the joint heatmaps. The part affinities are used as edge weights in the graph based clustering. In [5] pixelwise offset values are calculated pointing from one joint to another. These offsets are used for grouping. [6] proposes a graph convolutional network for clustering. Thus, the clustering can be learned as part of an end-to-end framework. As stated in [11] and [12], invisible joints and the small local context used for joint estimation lead to an inferior performance of bottom-up methods.

Top-down approaches first detect all persons within an image and then estimate the pose for every instance. Most works follow the heatmap based approach of [30]. Mask R-CNN [7] learns both stages in a single end-to-end trainable network. Recent top-down methods profit from better person detectors or better network architectures [8]–[10]. Still, dense person group situations remain challenging for top-down methods. On the one hand, estimating positions of occluded joints is difficult. On the other hand, image crops of detected persons contain parts of other persons as well. In some cases, the overlapping region between two persons is so high, that the target pose is ambiguous. [31] proposes a solution for the handling of occluded joints training separate heatmap estimators for occluded and visible joints. Thus they train different experts for different occlusion states but not for disambiguation of multiple persons within a crop. [13] tries to solve the ambiguity for multiple persons by adding the position of a visible joint point for each person as an additional input. They depend on the results of a state of the art bottom-up pose estimation approach for these input hints. In AlphaPose+ [11] detections are handled independently within the single person pose estimation. A so called joint candidate loss allows the estimation of all joints that are within an image crop. The disambiguation of poses of different persons is part of a post-processing stage. There, joint candidates from all heatmap estimations are extracted. In a global graph based optimization procedure they can be reassigned to different detections based on the heatmap scores. As it is a fixed algorithm it can not be trained end-to-end within the framework. [12] depends on initial pose results of AlphaPose+ [11], that are refined by a graph convolutional network (GNN) depending on image features

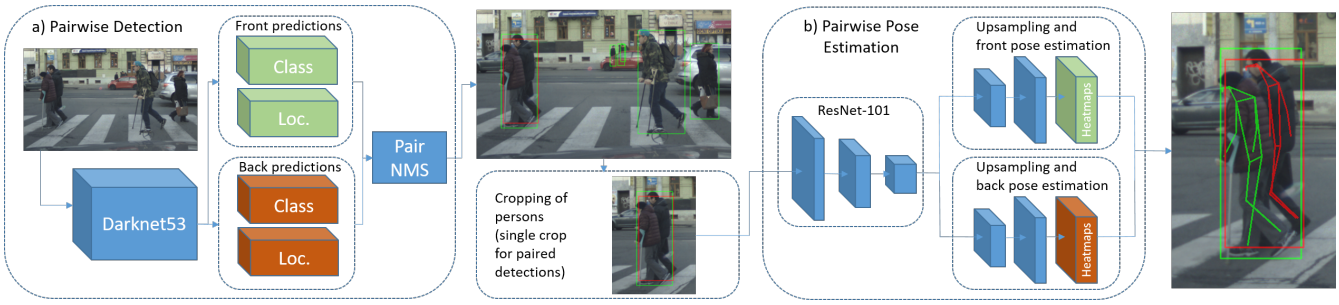

Fig. 2: Overview of our Simple Pair Pose method consisting of a pairwise detection a) and pose estimation method b).

extracted from the base network of AlphaPose+. They also propose a variant of this GNN, where poses of pedestrian pairs are jointly refined.

c) Datasets: Progress in deep learning based pose estimation has been driven by datasets like MSCOCO [16], MPII [17] and AI Challenger [18]. The CrowdPose [11] and OCPose [12] datasets focus on crowd situations with a high amount of pedestrians overlapping each other. These situations constitute a specific challenge for pose estimation. The images of the mentioned datasets have been collected using online search engines, Flickr and YouTube. In terms of automotive datasets, PedX [15] provides stereo images recorded from a moving vehicle including LiDAR annotated with 2D and 3D poses. Still, the diversity regarding context is rather low as only three urban intersections are covered. Recently, the TDUP dataset [14] has been announced, that will provide images recorded from a moving vehicle covering diverse urban traffic scenes in China. An overview of these datasets is shown in Table I.

d) Contributions: Our contributions are twofold. First, we propose a new top-down pose estimation method to jointly estimate poses of pedestrian pairs in a single network. It relies on paired detections using an adopted set detection approach [24] that improves the recall in groups. Our new method is simple to integrate in existing network architectures for human pose estimation, yet effective and does not depend on a separate input hint or a post-processing stage for disambiguation of poses of pedestrian pairs. Second, we provide a new automotive dataset for pose estimation extending the original ECP dataset by additional images from the front-facing and two side-facing cameras. The detailed annotations including bounding boxes and poses of pedestrians and riders will be made available on our website[b]. The annotations of the test dataset are kept private for fair benchmarking using our server, that will be extended by an automatic evaluation protocol.

III. Method

Our *Simple Pair Pose* (SPP) Method for top-down pose estimation consists of two parts (see Figure 2). For the *pairwise detection*, we integrate the idea of set detection of [24] into our YOLOv3 [25] detector to improve the recall in groups. In the *pairwise pose estimation* part, we extend the single person pose estimation network described in [11] to jointly estimate the poses of paired detections.

[b]https://eurocity-dataset.tudelft.nl

A. Set Detection Revisited

Deep learning based detection approaches like [25]–[28] depend on proposals as input. This raises two issues regarding detection in crowds. First, for a single pedestrian there are usually several overlapping proposals. The pedestrian is used as training target for all proposals, that are associated e.g. based on an IoU threshold. During inference, this results in multiple detections per pedestrian, that have to be suppressed by the NMS. Depending on the IoU threshold of the NMS, not all pedestrians within a crowd may be detected.

Second, within a group scenario, a single proposal often overlaps with several pedestrians. Still, many approaches only select a single person with the highest overlap as target for every proposal. In inference this may result in some kind of ambiguity. When a proposal is placed between two pedestrians the final detection may be influenced by both pedestrians and has a low localization accuracy [29]. To solve this issue [24] proposes to predict all objects associated with a single proposal. Therefore, the predictor head of a feature pyramid network [32] consisting of a classification and localisation part is duplicated. During training, all predictions from a single proposal are matched with the associated ground truth annotations minimizing an earth mover distance loss [24]. They also propose a *Set NMS* at which predictions from the same proposal do not suppress each other. This solves the first issue of missing detections within a crowd. We adopt the idea of [24] in our pairwise detection method.

B. Pairwise detection

YOLOv3 predicts bounding boxes based on three features layers of its Darknet-53 core network downscaled by a factor of 8, 16 and 32. Prior boxes of different sizes and aspect ratios are centered in every cell of these layers and serve as proposals. For every prior box, the prediction head estimates four coordinate offsets in the localisation part and the confidences for the different classes. We configure the NMS with an IoU threshold of 0.5 in accordance with [20], resulting in a low recall for pedestrian pairs. In our YOLOv3 extension we apply the idea of set prediction of [24] and focus on pedestrian pairs, as shown in part a) of Figure 2. Therefore, we assign a set cardinality of two. This reduction enables us to formulate the set prediction more explicitly. For pairs, we define the pedestrian with the lower bounding box edge to be the *front pedestrian*, whereas the other one is the *back pedestrian*. Following a flat world assumption this corresponds to the z-ordering in the traffic scene. We

manually annotate the ordering for pedestrians with an equal lower bounding box edge or contradicting occlusion levels.

We duplicate the prediction head of YOLOv3. For every prior box two predictions including two bounding box regressions and classifications are estimated. For the first prediction head we always set the front pedestrian as target, while the second prediction head is responsible for estimating the back pedestrian. We do not permutate the matching as in [24]. Thus, we train separate experts for both cases and disambiguate the detection task for pedestrian pairs, as it is defined beforehand which pedestrian has to be detected by which head. In [24] this has to be learned implicitly.

We model the bounding box regression loss ($\mathcal{L}_{loc}$) to follow a normal distribution as in [33] and the classification to follow a softmax loss ($\mathcal{L}_{cls}$), which enables uncertainty weighting [34]. We receive the following total loss for a prior box associated with a pedestrian pair,

$$\begin{aligned}\mathcal{L}(w) =& \mathcal{L}^f_{cls}(gt^f, w) + \mathcal{L}^f_{loc}(gt^f, w) \\ &+ \mathcal{L}^s_{cls}(gt^b, w) + \mathcal{L}^s_{loc}(gt^b, w)\end{aligned} \tag{1}$$

with gt^f, gt^b as the ground truth annotations of the front and back pedestrian, w as the weights of the network, and $\mathcal{L}^f, \mathcal{L}^s$ as the losses of the first and second prediction head.

If a pedestrian is not part of a pair, we define it to be a front pedestrian by default. In this case the regression loss for the second prediction head is zero, and its target class is background. For inference, similar to [24] we adapt the NMS in a way, that front pedestrians do not suppress back pedestrians estimated based on the same proposal (*Pair NMS* in Figure 2). If both class confidences of a front and back prediction from the same proposal are above a certain threshold, we define this as a *paired detection*.

C. Pairwise Pose Estimation

We follow the top-down multi person pose estimation approach: In general, detections are cropped from the input image and a single person pose estimation (SPPE) network estimates the n heatmaps of the n joints. If a crop contains several pedestrians it may be ambiguous which pose has to be estimated [13]. This is also caused by imperfectly localized detection boxes.

To avoid confusions of front and back joints of the front and back pedestrian, we jointly estimate heatmaps for both pedestrians in a single forward pass. In [11] a pose head consisting of two upsampling modules and a final convolutional layer is attached to a ResNet-101 backbone to estimate the pose heatmaps. We duplicate this pose head to jointly estimate front and back heatmaps as shown in part b) of Figure 2. It is possible to split the paths later (or even earlier) within the network, e.g. by only duplicating the final layer. The point to branch may be empirically selected, while an earlier split increases runtime.

During training, ground truth boxes gt^f and gt^b of pedestrian pairs are combined to a single pair box gt^p enclosing the two boxes. This combined box is used to crop the image to ensure that the context of both pedestrians is fully available. The overall heatmap loss is the sum of the separate heatmap losses for the joints of the front and the back pedestrian. By training separate experts for estimating the heatmaps of front and back pedestrians, we disambiguate the target pose. As before in the detection method, we define single pedestrians to be front pedestrians by default and use the single box for cropping the image. The heatmap loss for the back joints is zero in this case. During inference, paired detections of our pairwise detector are combined as in the training before cropping, while single detections are kept as they are. We extract the best front and back poses from the heatmaps using spatial argmax. Hence, we do not depend on a post-processing step to handle poses of pedestrian pairs. As all computations are shared apart from the duplicated pose head, the runtime for pedestrian pairs is lower in comparison with estimating the two poses based on separate image crops.

IV. Dataset

A. Data selection

The ECP dataset has been recorded from a moving vehicle in 31 cities of 12 European countries [20]. The two megapixels camera (1920x1080) attached behind the windshield has been operated with 20 frames per second for a total recording time of 60 hours. For the detection benchmark a fixed sample rate has been used to extract and annotate images to avoid any selection bias.

For our EuroCity Persons Dense Pose dataset we shift the main focus to crowded scenes. In addition to the front facing camera, two side facing cameras with a higher horizontal field of view of 85° had been attached at the left and right door mirrors. They feature the same resolution and have been synchronously triggered with the front facing camera.

We select images with a high number of persons from the front as well as side facing cameras. For images already contained in ECP this is done based on the number of box annotations. For the remaining images we run a Faster R-CNN [27] model trained on ECP to detect crowded scenes.

Overall, we select 30,704 images from the front facing camera, of which 14,438 are already part of the ECP dataset. Further 8,263 images from the left and 8,008 images from the right camera are added to the final image set consisting of 46,975 images in total. We ensure that the train-val-test split of our new ECPDP dataset is aligned with the train-val-test split of ECP [20].

B. Dataset Annotation

Apart from poses the annotation protocol mimics that of ECP [20]. For every image all pedestrians and riders of at least 20 pixels in height are annotated with tight bounding boxes of the complete extent. If a person is not fully visible, the extent is estimated. In that case, the level of occlusion and truncation is annotated. Groups of persons that are not distinguishable are annotated with boxes enclosing the groups, serving as ignore regions during evaluation. Ignore regions are also class-specific for pedestrians and riders. If the class can not be discriminated by the labeler, it is annotated as generic person ignore region. In addition, we annotate the complete poses consisting of 17 joint points as

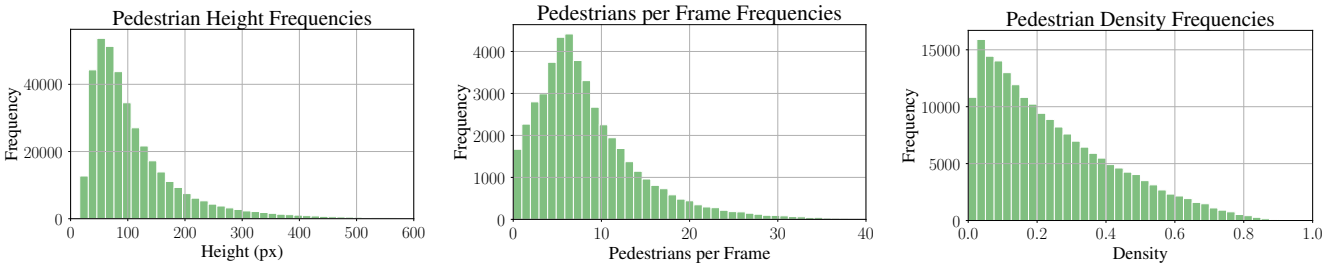

Fig. 3: Frequencies of our EuroCity Persons Dense Pose (ECPDP) dataset. The density of pedestrians (right) is only shown for overlapping boxes, meaning densities greater zero.

in MSCOCO [16] for persons that are greater than 60 pixels in height. For every joint, it is indicated if it is fully visible, self-occluded or occluded.

C. Dataset Statistics

Data distributions for pedestrians of our new dataset are shown in Figure 3. Due to our data selection targeted on crowded scenes, there is a peak around six pedestrians per frame. We analyse the overlap between pedestrians, as mutual occlusions of pedestrians cause major challenges even for recent deep learning approaches. As in [22], we define the density of a pedestrian as the highest IoU with any other pedestrian in this scene. As defined before, if the density is greater than 0.5, the two pedestrians form a pair. The amount of pairs in our ECPDP dataset is about one percentage point higher than in the ECP dataset (5.9% in contrast to 5.0%). Regarding riders, only 1.2% of these have a mutual IoU greater 0.5. Therefore, we focus on pedestrians only in our pairwise experiments.

Compared to other automotive datasets, our new ECPDP dataset provides the largest number of pose annotated persons (cf. Table I). Furthermore, the ECPDP contains the largest average number of persons per image overall. Thus, it enables the targeted evaluation of pose estimation in dense urban traffic scenes. Detailed statistics of the dataset subsets are shown in Table II.

TABLE II: Statistics of the subsets of our new ECPDP dataset regarding the number of images and the amount of boxes, poses and ignore regions of pedestrians and riders and the number of generic person ignore regions that may contain pedestrians as well as riders.

	train	val	test	**total**
# images	29,570	5,150	12,255	**46,975**
# pedestrian boxes	251,654	47,530	99,529	**398,713**
# pedestrian poses	167,066	30,960	65,698	**263,724**
# pedestrian ignore	17,140	3,394	7,255	**27,789**
# rider boxes	21,617	3,624	8,458	**33,699**
# rider poses	10,164	1,704	3,737	**15,605**
# rider ignore	943	150	347	**1,440**
# person ignore	9,158	1,783	3,605	**14,546**

D. Metrics

For evaluation of the detection performance, we apply the log average miss rate (LAMR) as in [20]. In many pair situations one of the two pedestrians has a rather low while the other has a high level of occlusion. Hence, pedestrians of a pair would be divided into the *reasonable* and the *occluded* subsets as defined in ECP. To have a common subset for pairs we add another subset named *relevant*. It consists of all pedestrians of at least 40 pixels in height and less than 80% occlusion.

We use the object keypoint similarity (OKS) from [16] to evaluate pose estimation accuracy. In our pairwise pose evaluation we match objects based on their IoU and measure the average OKS for true positives.

In [16], objects are matched based on their OKS instead of the IoU, as not all of the bottom-up methods provide bounding boxes. They calculate the average precision (AP) for different OKS matching thresholds. We apply the same evaluation procedure for the overall pose estimation performance that serves as baseline for benchmarking on our new pose dataset. Instead of calculating the AP we use the LAMR. We adapt the LAMR implementation of [20] by matching objects based on their OKS instead of the IoU. Samples without pose annotations or that are not part of an evaluation subset serve as ignore instances and are still matched based on the IoU if there is no other non-ignore instance that exceeds the OKS threshold for matching.

V. Experiments

We first focus on training and evaluation of our pairwise detection method for pedestrians. Then, we describe our training setup for the pairwise pose estimation, and show results of the pose estimation for pedestrian pairs. Finally we show the overall pose estimation performance on the complete ECPDP test dataset that serves as baseline on our new pose benchmark. We also include riders in the training of all our models. Still, as rider pairs are rare, we only train the front prediction head and the front pose heatmaps with riders and focus on pedestrians in the evaluation.

A. Pairwise detection training

We adopt the YOLOv3 tensorflow implementation of [33]. The nine prior box sizes are optimized on the ECP training dataset as in [20]. Flipping and crop and scale augmentation is used for all trainings.

We first train a model on the ECP day-time training subset with the default prediction head of YOLOv3, that estimates a single detection per prior box. The Darknet-53 part of this

Fig. 4: Qualitative results of AlphaPose+ (left) and our pairwise pose estimation (right) for back pedestrians (red) and front pedestrians (green) of valid paired detections (green and red bounding boxes). The first two rows show samples where our method surpasses AlphaPose+, while the last row shows error cases.

network is initialized with weights trained on ImageNet for classification [35]. We train for 33 epochs in total, decreasing the initial learning rate of 1e-5 after 13 and 25 epochs by a factor of 0.1. The best performing checkpoint is selected on the ECP validation dataset and serves as *Init* model.

For our *Base* model the training of the *Init* model is continued on the training subset of our new ECPDP dataset. It is trained for 50 epochs, reducing the initial learning rate of 1e-5 after 30 and 44 epochs by a factor of 0.1.

Finally, we train our pairwise detection network also on ECPDP. The *Init* model is used for initialization. The weights of the additional convolutional filters of the second prediction head for estimating the classification and bounding box regression of the back pedestrian are randomly initialized. We achieved best results with a fixed weighting for the losses of the second prediction head instead of uncertainty weighting [34] that is used for the losses of the first prediction head. The *Pair* model is also trained for 50 epochs with the same learning rate strategy as the *Base* model.

B. Pairwise detection results

The detection performance for pedestrians is evaluated on the *relevant* subset of our ECPDP test dataset.

We evaluate a version of the *Pair* model discarding all back predictions (coined *Pair w/o back*). This removes the influence of back predictions, that on the one hand increase the recall in groups but on the other hand decrease precision due to false positives. A greedy NMS with an IoU threshold of 0.5 is applied for this version of the *Pair* model and the *Base* model. The full *Pair* model including back predictions makes use of our adapted NMS as described in Section III-B.

Quantitative results are shown in Table III. The LAMR of the two *Pair* model variants is 0.8 points higher than of the *Base* model. The additional prediction head slightly reduces the overall detection performance. Still, the recall for pedestrian pairs can be increased by the back predictions. Despite the NMS threshold of 0.5, the recall of the *Base* model for pairs at a false positive per image (fppi) rate of 1.0 is also greater 50%. This is caused by imperfectly localized predictions that are not suppressed by the greedy NMS.

Results of valid paired detections are shown in Figure 4. In Figure 5 the recall is shown for different density ranges of the test samples for a fppi rate of 1.0. The *Pair* model achieves the highest recall for density ranges above 0.5.

TABLE III: Detection results for pedestrians of the *relevant* subset on the ECPDP test subset. All values are given in percentage points. $\text{Rec}_{>0.5f:x}$ is the recall for pedestrians of pairs with a mutual IoU greater 0.5 for a given false positive per image (fppi) rate of x.

Model	LAMR	$\text{Rec}_{>0.5f:0.1}$	$\text{Rec}_{>0.5f:1}$
Base	**28.2**	51.7	62.1
Pair w/o back	29.0	49.1	62.0
Pair	29.0	**55.5**	**70.0**

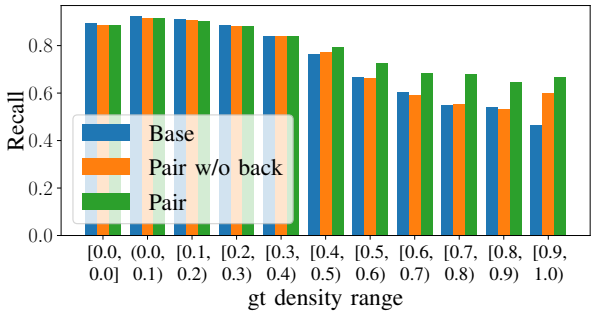

Fig. 5: Recall of pedestrians normalized per bin for our three detection models in dependence of the density of the test samples. The density is defined as the highest IoU with any other test sample.

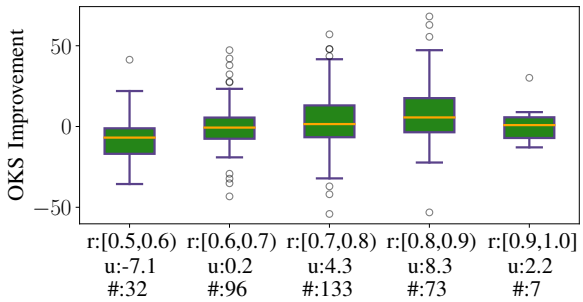

Fig. 6: Mean OKS^b_{all} improvement u for back pedestrians of our method in comparison with AlphaPose+ in dependence of the density of the paired predictions binned over different density ranges r with # as the number of samples per bin.

C. Pairwise pose training

For better comparability with AlphaPose+ [11] we use the provided source code[c] and integrate our pairwise pose estimation method. Therefore, we duplicate the pose head consisting of two upsampling modules and a final convolutional layer attached to a ResNet-101 [36] backbone, combine pedestrian pairs before cropping, and simply skip the graph based optimization of AlphaPose+. (Only duplicating the final convolutional layer has lead to inferior results in our experiments.) By using the framework of AlphaPose+, we verify the straightforward integrability into other methods. The joint candidate loss proposed in [11] is not provided in their framework. For training of the AlphaPose+ baseline and our pairwise pose estimation we use the mean squared error as heatmap loss. The training settings are identical for both methods. The person crops from the input image are rescaled to 320 x 256. The output heatmap resolution is 80 x 64. We train for 110 epochs, reducing the initial learning rate of 1e-4 to 1e-5 after 80 epochs.

TABLE IV: Pose results in terms of mean OKS (median in brackets) for detected pedestrian pairs of the ECPDP test subset. All values are given in percentage points.

Model	OKS^f_{vis}	OKS^f_{all}	OKS^b_{vis}	OKS^b_{all}
AlphaPose+	85.6 (93.3)	83.8 (88.9)	68.7 (75.8)	65.5 (68.7)
SPP (ours)	**86.9** (**94.3**)	**84.9** (**89.7**)	**75.9** (**81.9**)	**68.3** (**71.7**)

D. Pairwise pose results

For evaluation of our *SPP* method on the test dataset, we run the pairwise pose model on the detections of our *Pair* detection model including back predictions. Paired pedestrian predictions are combined first to jointly estimate the front and back pose. The predictions of the *Pair* model are also used as input for AlphaPose+ for better comparability and as YOLOv3 is also the underlying detection method in [11].

We focus on evaluation of the pair scenarios and show mean and median OKS values on the 346 correctly detected pedestrian pairs in Table IV. The two estimated poses are associated with the front and back ground truth poses optimizing the overall OKS value. OKS^f and OKS^b are the OKS values for front and back pedestrians. All joints or only visible joints are taken into account for OKS_{all} and OKS_{vis}. Our model performs best for front as well as back pedestrians of pairs. Most significant improvement can be observed for back pedestrians, which is 7.2 percentage points for the mean OKS evaluated on visible joint points (OKS^b_{vis}) and 2.8 points on all joints (OKS^b_{all}). AlphaPose+ performs similarly for poses of front pedestrians. In the qualitative results in Figure 4 there are several cases where AlphaPose+ confuses poses of the front with the back pedestrians. This is caused by missing joint candidates for the pedestrians in the back, whereas our model profits from the expert knowledge for the back pedestrians. We show the OKS improvement of our method in comparison with AlphaPose+ for these back pedestrians in dependence of the IoU between the paired detections in Figure 6. Apart from the last bin, that only contains seven samples, a higher IoU between the detections results in a higher average improvement by our method. This can be expected as a higher overlap between detections may also induce more difficulties in discriminating the two pedestrians within the pose estimation. This higher overlap can be also caused by a low localization accuracy of the pair detector, e.g. when the pair detector itself confuses extents of front and back pedestrians. Our method suffers less from these localization errors of the underlying detector as the two boxes are combined and the disambiguation is solved by the different experts.

TABLE V: Overall pose performance on the ECPDP test subset. All values are given in percentage points. For the LAMR L^t_o test samples occluded up to o% are matched based on an OKS threshold t. Pairwise training is only applied for pedestrians in our SPP method.

Model	Class	Scores	$L^{0.5}_{40}$	$L^{0.75}_{40}$	$L^{0.5}_{80}$	$L^{0.75}_{80}$
AlphaPose+	Ped.	Box	33.9	56.7	41.1	64.0
AlphaPose+	Ped.	Pose	29.8	49.3	36.1	56.2
SPP (ours)	Ped.	Box	32.0	56.3	39.8	63.8
SPP (ours)	Ped.	Pose	**28.9**	**48.8**	**35.9**	**56.0**
AlphaPose+	Rider	Pose	**11.2**	**19.0**	**13.7**	**23.1**
SPP (ours)	Rider	Pose	11.5	**19.0**	14.1	**23.1**

[c]https://github.com/MVIG-SJTU/AlphaPose/tree/pytorch

E. Overall pose results

For benchmarking purposes on our new dataset, we show the OKS based LAMR (abbreviated as L in the following) on all test samples annotated with poses for pedestrians and riders in Table V. We use two different OKS thresholds for matching: 0.5 for $L^{0.5}$ and 0.75 for $L^{0.75}$ respectively. L_{40} is calculated for persons less than 40% occluded and L_{80} for less than 80% occlusion. As before, the detections from the *Pair* model are used for inference of AlphaPose+ and our pairwise pose model. For pedestrians, we compare results using the confidences from the box detector and the confidences from the pose estimation, where heatmap scores are added to the initial class scores. Confidences from the pose estimation result in better performance for both models. In our methodology we focus on pedestrian pair situations. As the amount of pedestrian pair situations is low in comparison with all test samples the overall performance is only slightly better than for AlphaPose+, e.g. by 0.9 points for $L_{40}^{0.5}$. The performance metric for pedestrians up to 80% occlusion is similar. Note that we do not make use of pairwise training for riders in our *SPP* method due to the low relative amount of pairwise rider situations. Therefore, the results for riders in Table V for our model does not show any improvement over AlphaPose+.

VI. Conclusion

In this work we presented our new *Simple Pair Pose* method for top-down human pose estimation. The underlying YOLOv3 detector extended by the set detection idea of [24] improves the recall in groups by jointly detecting pairs of pedestrians. We have shown experimental results for our new pose estimation method, that jointly predicts poses for both pedestrians of these pairs. As all computations are shared apart from the final duplicated layers, it reduces the runtime for paired detections in comparison with separate pose estimation. Yet, by implicitly training different experts for poses of front and back pedestrians, it is very effective and surpasses the AlphaPose+ method used for comparison. Our approach could be easily integrated in other heatmap based single person pose estimation approaches. It could also be used as input for the recent graph based method of [12] that relies on input poses from AlphaPose+. The combination with other detection methods optimized for crowded scenes is left for future work. Furthermore, we presented our new EuroCity Persons Dense Pose dataset, which will serve for benchmarking of pose estimation methods on dense urban scenarios. By releasing our dataset and providing automatic evaluation scripts on our server, we hope to further enable research in this exciting field.

References

[1] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *Proc. IEEE WACV*, 2021, pp. 1258–1268.

[2] L. Pishchulin *et al.*, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE CVPR*, 2016, pp. 4929–4937.

[3] E. Insafutdinov *et al.*, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. IEEE ECCV*, 2016, pp. 34–50.

[4] Z. Cao *et al.*, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. IEEE CVPR*, 2017, pp. 7291–7299.

[5] G. Papandreou *et al.*, "PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Proc. IEEE ECCV*, 2018, pp. 269–286.

[6] S. Jin *et al.*, "Differentiable hierarchical graph grouping for multi-person pose estimation," in *Proc. IEEE ECCV*. Springer, 2020, pp. 718–734.

[7] K. He *et al.*, "Mask R-CNN," in *Proc. IEEE ICCV*, 2017, pp. 2961–2969.

[8] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. IEEE ECCV*, 2016, pp. 483–499.

[9] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. IEEE ECCV*, 2018, pp. 466–481.

[10] K. Sun *et al.*, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE CVPR*, 2019, pp. 5693–5703.

[11] J. Li *et al.*, "CrowdPose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE CVPR*, 2019, pp. 10 863–10 872.

[12] L. Qiu *et al.*, "Peeking into occluded joints: A novel framework for crowd pose estimation," *Proc. IEEE ECCV*, 2020.

[13] S. Hong *et al.*, "HintPose," *arXiv:2003.02170*, 2020.

[14] S. Wang *et al.*, "UrbanPose: A new benchmark for VRU pose estimation in urban traffic scenes," in *IEEE IV*, 2021 (accepted for publication).

[15] W. Kim *et al.*, "PedX: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections," *IEEE RA-L*, vol. 4, no. 2, pp. 1940–1947, 2019.

[16] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. IEEE ECCV*, 2014, pp. 740–755.

[17] M. Andriluka *et al.*, "2d human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE CVPR*, 2014.

[18] J. Wu *et al.*, "AI Challenger: A large-scale dataset for going deeper in image understanding," *arXiv:1711.06475*, 2017.

[19] N. Bodla *et al.*, "Soft-NMS – improving object detection with one line of code," in *Proc. IEEE ICCV*, 2017.

[20] M. Braun *et al.*, "EuroCity Persons: A novel benchmark for person detection in traffic scenes," *IEEE TPAMI*, vol. 41, no. 8, pp. 1844–1861, 2019.

[21] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE CVPR*, 2017, pp. 4507–4515.

[22] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE CVPR*, 2019, pp. 6459–6468.

[23] J. Zhang *et al.*, "Attribute-aware pedestrian detection in a crowd," *arXiv:1910.09188*, 2019.

[24] X. Chu *et al.*, "Detection in crowded scenes: One proposal, multiple predictions," in *Proc. IEEE CVPR*, 2020, pp. 12 214–12 223.

[25] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv:1804.02767*, 2018.

[26] S. Ren *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. in NIPS*, 2015, pp. 91–99.

[27] R. B. Girshick, "Fast R-CNN," in *Proc. IEEE ICCV*, 2015, pp. 1440–1448.

[28] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. IEEE ECCV*, 2016, pp. 21–37.

[29] X. Wang *et al.*, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE CVPR*, 2018, pp. 7774–7783.

[30] J. Tompson *et al.*, "Efficient object localization using convolutional networks," in *Proc. IEEE CVPR*, 2015, pp. 648–656.

[31] T. Golda *et al.*, "Human pose estimation for real-world crowded scenarios," in *Proc. AVSS*, 2019, pp. 1–8.

[32] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE CVPR*, 2017, pp. 2117–2125.

[33] F. Kraus and K. Dietmayer, "Uncertainty estimation in one-stage object detection," in *Proc. IEEE ITSC*, 2019, pp. 53–60.

[34] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE CVPR*, 2018, pp. 7482–7491.

[35] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, 2009, pp. 248–255.

[36] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.