IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 February 2026, accepted 24 February 2026, date of publication 6 March 2026, date of current version 11 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3671610

RESEARCH ARTICLE

A Pipeline for Evaluation of Keypoint-Based Bounding Boxes for Multi-Scale Pedestrians

SARFRAZ AHMED[1], M. NAZMUL HUDA[2], CHITTA SAHA[3], AND MOHAMMED QUDDUS[4]

[1]Centre for Future Transport and Cities, Coventry University, CV1 5FB Coventry, U.K.
[2]Department of Electronic and Electrical Engineering, College of Engineering, Design and Physical Sciences (CEDPS), Brunel University of London, Uxbridge, UB8 3PH London, U.K.
[3]School of Engineering and Built Environment, Birmingham City University, B4 7XG Birmingham, U.K.
[4]Department of Civil and Environmental Engineering, Imperial College London, South Kensington Campus, SW7 2AZ London, U.K.

Corresponding author: M. Nazmul Huda (nazmul.huda@brunel.ac.uk)

This work was supported by Brunel University of London.

ABSTRACT Pedestrians account for approximately 25% of traffic accidents, many of which can be prevented using autonomous driving systems (ADS). Although pedestrian detection has advanced significantly, intent prediction still lags behind human perception. A key challenge is predicting the intent of smaller pedestrians, who are harder to detect and analyse using 2D pose estimation techniques because of their tendency to blend into the background. However, human joint keypoints (e.g., head, shoulders, elbows, and knees) can reliably predict pedestrian movement, such as gait, limb motion, and head orientation. This paper introduces the Keypoint Evaluation (KeyEval) pipeline, a new technique for generating high-quality pedestrian keypoints using a 2D pose estimator. Leveraging ground-truth bounding box annotations from the JAAD and PIE datasets, we assess keypoint accuracy and apply state-of-the-art fine-tuning, achieving a 19% improvement in average precision (76%) over the baseline. This suggests KeyEval can enhance predictions of pedestrian intent—crossing, waiting, or changing direction—particularly for smaller pedestrians. The KeyEval pipeline can be seamlessly integrated into ADS to proactively reduce vehicle-pedestrian accidents, as demonstrated in our previous work.

INDEX TERMS Autonomous driving system, intent prediction, pedestrian safety, pose estimation.

I. INTRODUCTION

In our previous study, [1], we introduced a state-of-the-art pedestrian intent predictor designed for multi-scale pedestrians. In that work, we used the JAAD [2] and PIE [3] datasets to generate high-quality keypoints. These keypoints were used to predict the crossing intentions of pedestrians in near real-time. In this study, we expand the process of generating high-quailty keypoints.

Human pose estimation aims to represent a person's orientation based on anatomical joints (e.g., neck, shoulders, wrists). These joints are commonly referred to as keypoints [4], [5]. Figure 1 illustrates the pose estimation for pedestrians performing various actions based on predicted keypoints. Recently, pose estimation has become integral to autonomous vehicle (AV) applications, particularly pedestrian intent prediction. These applications primarily focus on distinguishing between crossing and not crossing behaviours (e.g., [3], [6], [7], [8]), providing essential information for safe AV path planning and manoeuvring. Many of these studies rely on large-scale datasets, such as the Joint Attention in Autonomous Driving (JAAD) dataset [2], includes behavioural annotations (e.g., crossing, standing, stopping). More recently, the same authors introduced the Pedestrian Intention Estimation (PIE) dataset [3], an even larger dataset. However, neither dataset includes ground-truth keypoint annotations. Consequently, previous works have depended on off-the-shelf pose estimators for keypoint prediction. Despite this reliance, many studies do not evaluate the accuracy of their predicted keypoints or report the number of usable pedestrian samples. This omission affects the reproducibility and comparability of their approaches. Additionally, prior work primarily focuses on larger pedestrians, neglecting smaller pedestrians, who

The associate editor coordinating the review of this manuscript and approving it for publication was Tariq Umer.

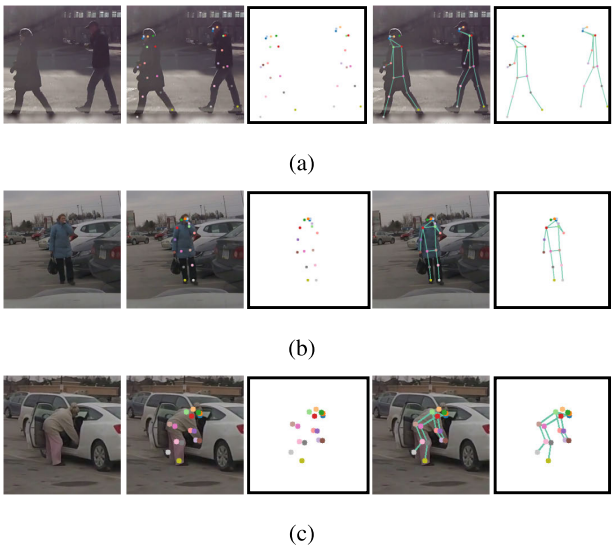


FIGURE 1. The joints predict the location of various anatomical points on the body. These joints are connected to create a skeleton of the pedestrian. Based on the skeleton and relative angles of the various points, a pose can be estimated such as walking (1a), standing (1b) and bending down (1c).

often appear at low resolution due to background blending and motion blur—making pose estimation more challenging. To address these limitations, this paper presents a novel approach for evaluating and improving predicted keypoints for multi-scale pedestrians in datasets without ground-truth keypoint annotations. Our approach specifically targets the JAAD and PIE datasets to enhance pose estimation for pedestrians of varying sizes. As mentioned, these datasets include crossing behavioural annotations, which were used for real-time pedestrian crossing intention prediction.

In this work, we introduce the Keypoint Evaluation (KeyEval) pipeline, a method for generating high-quality keypoints. Unlike previous pedestrian intent prediction approaches that rely on off-the-shelf pose estimators, KeyEval evaluates the accuracy of predicted keypoints. Using the most accurate keypoints, the pose estimator is then fine-tuned to enhance its performance. To the best of our knowledge, KeyEval is a novel approach for addressing dataset limitations related to missing ground-truth keypoint annotations. We have demonstrated the effectiveness of using the keypoints predicted using the KeyEval pipeline in our previous works (see [1], [9]). The purpose of this paper is to discuss the the KeyEval pipeline in further detail.

II. RELATED WORKS

A real-time context-invariant technique for predicting pedestrian crossing intentions was proposed in [7]. The architecture consisted of a multi-branch network which uses Cartesian features and location-invariant geometric skeleton features to predict crossing intentions. According to the authors in [10], the posture of a pedestrian is a vital aspect of an intention prediction model. The proposed approach achieving 94.4%, an improvement of nearly 6% when compared to the next best approach. However, it is mentioned that using an off-the-shelf detector in this manner is a drawback.

According to [6], using keypoints generated by 2D pose estimation is a promising avenue for predicting pedestrian intent prediction. Their proposed technique consisted of a stand-alone detector, off-the-shelf pose estimator and random forest for classification. Using this architecture, the authors achieved an accuracy of 88% for crossing/not crossing behaviours. Only pedestrians of more than a width of 50 pixels were considered. Again, this technique does not mention the evaluation techniques applied to determine the accuracy of the keypoints generated by the pre-trained pose estimator.

In [11], the authors propose a technique which utilises the relationship between pedestrians and objects within their surrounding environment. Using rich visual features and graph convolutional autoencoders the authors were able to achieve an F1-score of 79% for predicting a pedestrian's crossing intentions on the PIE dataset. The architecture uses an off-the-shelf pose estimator [4] to generate keypoints. This approach was able to extract over 740,000 unique poses for the 1,800 pedestrians in the dataset.

To overcome the obstacles and drawbacks of previous the previous techniques and methods, in [1], we proposed an approach which utilises not only utilises a pose estimator designed for multi-scale pedestrian keypoint, but also a method of fine-tuning the pose estimator to improve the quality of predicted keypoints for both the JAAD and PIE datasets. We demonstrated the effectiveness of the high-quality keypoints and useful information, such as posture and head orientation that they provide. In [1], the predicted keypoints over time (i.e., along frames) were passed through a LSTM-based model to predict whether a pedestrian would cross or not with respect to the vehicle. This approach outperformed previous state-of-the-art techniques by up to 7%, achieving an accuracy of up to 94%, while also maintaining a comparable run-time of 6.1ms.

III. METHODOLOGY

The proposed pipeline consists of three distinct components: pose estimator, keypoint evaluator and fine-tuner. The flowchart for the pipeline is illustrated in Fig. 2. The images (i.e., data) from the dataset are fed into the pose estimator. The pose estimator predicts keypoints for the pedestrians in the image. Based on the predicted keypoints, bounding boxes are generated. The process of generating bounding boxes is described in detail in Section IV-C. As previously mentioned, the JAAD and PIE datasets do not include keypoint annotations. Therefore, to evaluate the predicted keypoints, the generated bounding boxes are utilised. These generated bounding boxes are compared with ground-truth bounding boxes from the dataset for evaluation. During the evaluation phase, each generated bounding box receives a confidence score (C), which indicates the accuracy of the generated bounding box with respect to the ground-truth

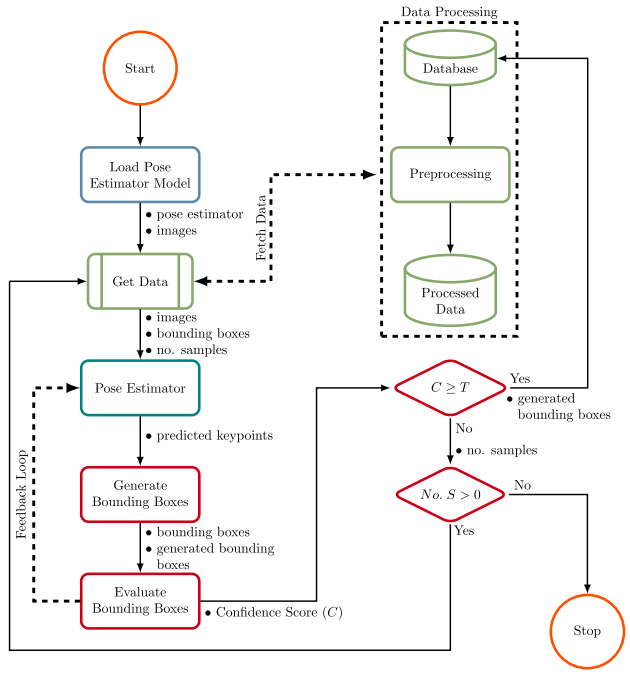


FIGURE 2. KeyEval pipeline flowchart.

bounding box. The associated keypoints for the bounding boxes with a specific confidence score (e.g., $\geq$ 50%) are added to the dataset while the generated bounding box replace the ground-truth bounding box. The keypoints are used for fine-tuning the pose estimator. Unlike previous techniques (e.g., [6]), the KeyEval pipeline is a multi-task approach, which can predict keypoints for multiple pedestrians in an image.

A. POSE ESTIMATION

For this task, the DEKR (disentangled keypoint regression) [5] approach was utilised to predict keypoints for multi-scale pedestrians. The DEKR pose estimator is a bottom-up pose estimator, meaning that a dedicated detector is not required for predicting pedestrian keypoints. As speed is a crucial aspect in real-time AV applications, this approach reduces the run-time to predict keypoints. The DEKR approach also outperforms previous benchmark datasets. For the COCO dataset, an AP of 70% was achieved, an improvement of 1.6% from the next best technique [5]. For the CrowdPose dataset, better results were reached, with a 2.6% improvement over previous techniques with an AP of 68%. Although other similar techniques exist, such as RTMPose [12] and HigherhrNet [13], the DEKR model is highly effective in complex scenarios, such as crowded streets [14].

B. KEYPOINT EVALUATION

As will be discussed in Section IV-C, bounding box coordinates can be calculated using the predicted keypoints. As the JAAD and PIE datasets do include ground-truth bounding boxes, the generated bounding boxes can be evaluated. The evaluation compares the generated and ground-truth

TABLE 1. Multi-scale settings.

Setting	Height (h)	Occlusion (o)
Reasonable	$h \geq 150$	$o \leq 75\%$
Reasonable (small)	$60 \leq h < 150$	$o \leq 75\%$
Heavy Occlusion	$h \geq 150$	$o > 75\%$
All	$h \geq 60$	$o \leq 75\%$

Height refers to height in terms of pixels. Occlusion is the level of visibility.

bounding boxes. Fig. 3b illustrate how the keypoints are used for generating keypoint-based bounding boxes (this process is discussed in Section IV). The green boxes are ground-truth bounding boxes, representing large and smaller pedestrians, respectively, while the blue boxes are the keypoints-based generated bounding boxes. The skeleton is removed in Fig. 3c for a clearer view of the ground-truth and generated keypoints. For this example in Fig. 3c, any bounding boxes with a confidence score of over 50% are regarded as correct predictions (further discussed in Section IV). In Fig. 3d, the ground-truth boxes are replaced by the generated boxes. The dataset is also updated with the inclusion associated keypoints by which the generated bounding boxes are based upon. This, allows for fine-tuning the pose estimator used in the KeyEval pipeline as it provides high-confidence pseudo keypoints for the JAAD and PIE datasets.

C. FINE-TUNING

Although the DEKR pose estimator achieved state-of-the-art for the COCO and CrowdPose dataset benchmarks, it is worth noting that these datasets have limitations. One limitation is both datasets have persons as the focal point in the images. This is not naturalistic as pedestrians will not always be the focus of the image in real-world applications, particularly smaller pedestrians [15]. Also, datasets, such as JAAD and PIE datasets, are much more complex, with lots of other objects in the scenes and with varying lighting conditions. Some objects occlude or bear a similar resemblance to pedestrians, making it more difficult to accurately predict keypoints. For this reason, fine-tuning techniques were applied to improve the accuracy of predicted keypoints.

IV. EXPERIMENTATION

A. MULTI-SCALE PEDESTRIAN SETTINGS

The widely used multi-scale settings for pedestrian detection as found in [16] were employed in this paper. These settings include evaluation of pedestrians of varying heights and visibility (occlusion) (see Table 1). As the images in the JAAD and PIE dataset are approximately three times larger than the images used in [16], the multi-scale settings are adjusted accordingly. For example, in [16], a pedestrian with a height of 50 pixels or more is considered as a *reasonable* pedestrian. Therefore, in this work, we consider pedestrians with a height of 150 pixels or more to be *reasonable*.

The JAAD dataset consists of over 2,000 unique pedestrians with over 300,000 annotations. When omitting heavily occluded instances, there are 234 "crossing" samples and

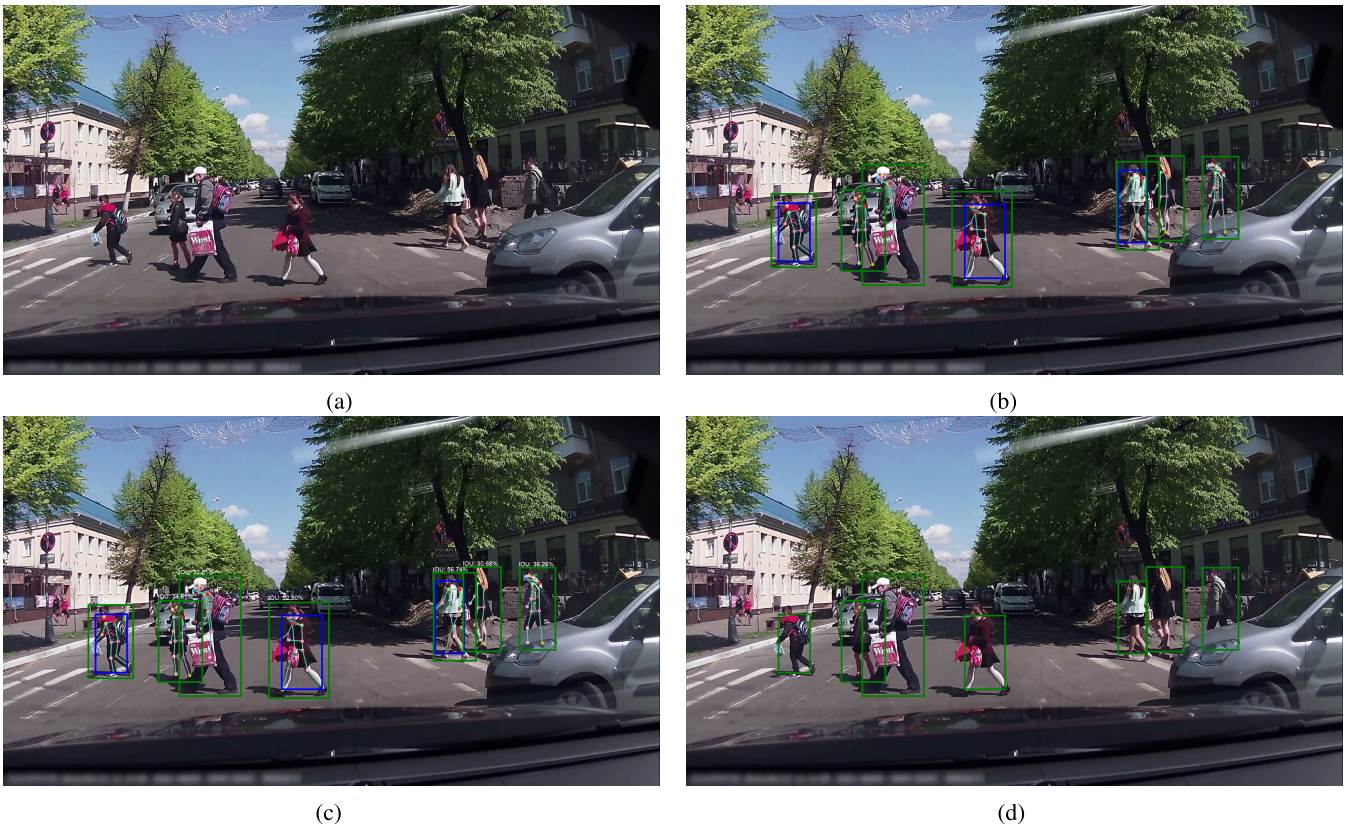
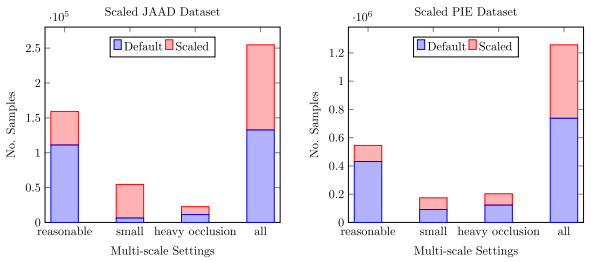


FIGURE 3. Generating Keypoint-based Bounding Boxes (Example 1): The input image (3a) is fed into the model, which predicts keypoints for the pedestrians in the image (3b). Based on the predicted keypoints, bounding boxes are calculated. The generated bounding boxes (blue rectangles) are compared with ground-truth bounding boxes (green rectangles) and a confidence score is calculated (3c). Based on the confidence scores, the ground-truth bounding boxes are replaced by generated bounding boxes (3d). In this example, the confidence score was set to 50%.

81 "not crossing" samples. The PIE dataset consists of approximately 1,800 unique pedestrians with over 700,000 annotations. There are 512 and 430 "crossing" and "not crossing" behaviour annotations respectively. A drawback of the JAAD dataset is that the crossing annotations are unbalanced, with significantly more "crossing" samples. The PIE dataset not only provides a larger number of samples, but also it provides more balanced crossing annotations. Both datasets consist of videos with a resolution of 1920 × 1080 at 30 fps (frames per second). There are significantly more"crossing" samples when compared to "not crossing" samples, with a ratio of approximately 1:6. The "irrelevant" class is ignored as these samples are not useful based on the dataset annotations provided by the authors. This type of data imbalance can lead to model classifiers being unable to generalise. This is caused by the model developing a bias towards the class with the larger number of samples. Balancing the data, therefore, aids in preventing such biases.

B. DATA AUGMENTATION

Both the JAAD and PIE datasets have a limited number of small pedestrian samples. Adding to that, localising smaller pedestrians is a challenging task due to the difficulty of distinguishing them from overlapping pedestrians and cluttered backgrounds [17]. This leads to blurred boundaries, which obscures their appearance. On the other hand, larger pedestrians generally exhibit specific characteristics, with rich information, which are features the model associates with those larger pedestrians. Therefore, data augmentation techniques, other than over-sampling were implemented. One approach was to predict keypoints for larger pedestrians and scale both the keypoints and images down by approximately 70%. This strategy provided further smaller-scale pedestrians. It was found that scaling the image down further did not reduce the image resolution too drastically, becoming unusable. The re-scaled images and keypoints were added to the original dataset, providing not only an increased number of smaller samples but also further samples of the *reasonable*, *heavy occlusion* and *all* categories (see Fig. 4, small represents the reasonable (small) stetting). Although these are re-scaled samples, they can be used as new samples as the decreased resolution results in the re-scaled samples being more challenging. Providing more challenges improved the ability of the model to generalise. In Fig. 4, for the JAAD dataset, this implementation increased small pedestrian samples by a magnitude of 7, while increasing the reasonable, heavily occluded and overall samples by 43%, 49% and 89% respectively. For the PIE dataset, the smaller



(a) (b)

FIGURE 4. Scaled Samples: Re-scaled images and keypoints to provide increased number of overall samples, focusing on improving the limited number of reasonable (small) samples. Reasonable (small) samples use the "small" label in this figure.

$$\begin{bmatrix} x_1 & y_1 & score \\ x_2 & y_2 & score \\ \vdots & \vdots & \vdots \\ x_n & y_n & score \end{bmatrix}$$

FIGURE 5. HigherHRNet Keypoint Predictions Format.

pedestrian samples were nearly doubled. The reasonable, heavily occluded and overall samples increased by 27%, 45% and 71%.

C. KEYPOINT-BASED BOUNDING BOX COORDINATES

As previously discussed, the JAAD and PIE datasets do not include ground-truth keypoint annotations. Because of this predicted keypoints cannot be evaluated. However, both datasets do include ground-truth bounding box annotations. Therefore, a method for using keypoint-based generated bounding boxes is proposed. As mentioned, the DEKR pose estimation model is pre-trained on the COCO and CrowdPose datasets. The predicted keypoints are in the format 17×3 and 14×3 for the COCO and CrowdPose datasets respectively. Values 17 and 14 refer to the keypoints or joints that each dataset focuses on, which are slightly different. For example, the COCO dataset has 3 more facial keypoints that are considered, the eyes, ears and nose, whereas the CrowdPose dataset only has the head. The first two rows represent the (x, y) coordinate of the predicted keypoint while the third row is the confidence score of each predicted keypoint (see Fig. 5). Any confidence score of less than 0.5 is omitted. This means only the highly accurate keypoints are used to generate the associated bounding boxes. Based on the (x, y) coordinates, associated bounding boxes can represent by (1). The values for $left, top, right, bottom$ is calculated via (2)-(5). Using the intersection-over-union (IoU) metric, the keypoint-based bounding boxes are evaluated. If the IoU value of these boxes is larger than the threshold value, then it is counted as a correct detection, whereas, if the value is lower than the threshold, it is considered a false detection.

$$box = [left, top, right, bottom] \tag{1}$$
$$left = \min([x_1, x_2 \ldots x_n]) \tag{2}$$
$$right = \max([x_1, x_2 \ldots x_n]) \tag{3}$$
$$top = \min([y_1, y_2 \ldots y_n]) \tag{4}$$
$$bottom = \max([y_1, y_2 \ldots y_n]) \tag{5}$$

By evaluating the generated bounding boxes, the accuracy of the predicted keypoints will also be calculated as the bounding boxes are based on the keypoints. Not only will this method be used for validating the accuracy of the predicted keypoints, but the most accurate keypoints were used for fine-tuning the DEKR pose estimation model for improving keypoint predictions. This will be discussed in the following sections.

D. TRAINING & EVALUATION PROTOCOLS

For training, the feedback process from Fig. 2 is illustrated in Fig. 6. As discussed in IV-C, the JAAD and PIE datasets do not consist of ground-truth keypoint annotations. Using the pre-trained DEKR pose estimator, keypoint-based bounding boxes are generated and evaluated. Based on their confidence score (C) is calculated. Confidence Score is determined by the IoU metric. If the value of C is greater then some threshold (T), the generated bounding boxes and predicted keypoints are added to the original dataset. If the sample from *Get Data* includes stored keypoint values, the store keypoints and predicted keypoints are compared and the loss is calculated. The stored keypoints would be generated in a previous iteration. Based on the loss value, the *Optimizer* adjusts the model parameters. Over time (i.e., over iterations), the model learns keypoints for pedestrians of the JAAD and PIE datasets. We used the training settings from the original paper for the DERK model (see [5]). We utilised the Adam optimiser with an initial learning rate of 1e-3. The learning rate was dropped to 1e-4 after the 30th iteration and further dropped to 1e-5 after the 60th iteration with a batch size of 12 images. This approach is not computationally expensive, as this technique can prcoess a single image at a time. However, this also increases the overall processing time of the training stage.

V. RESULTS & DISCUSSIONS

A. EVALUATION METRIC

In the following sections, we discuss our findings in terms of accuracy. This section discuses how we calculate our results and justifies our evaluation metric. For the evaluation of the keypoint-based bounding boxes, the mean average precision (mAP) metric was utilised. The mAP is a popular and widely used metric for pedestrian detection and poses estimation (e.g., [5], [13], [17], [18]), where predicted bounding boxes are evaluated by comparing them to ground-truth bounding boxes. In the same way, the keypoint-based bounding boxes were compared with the ground-truth bounding boxes. The mAP metric utilises the Jaccard Index to measure the overlap of the ground truth and generated bounding boxes. Jaccard Index is represented by (6), where X is the ground-truth bounding boxes and Y predicted bounding boxes. In applications, such as pedestrian detection and

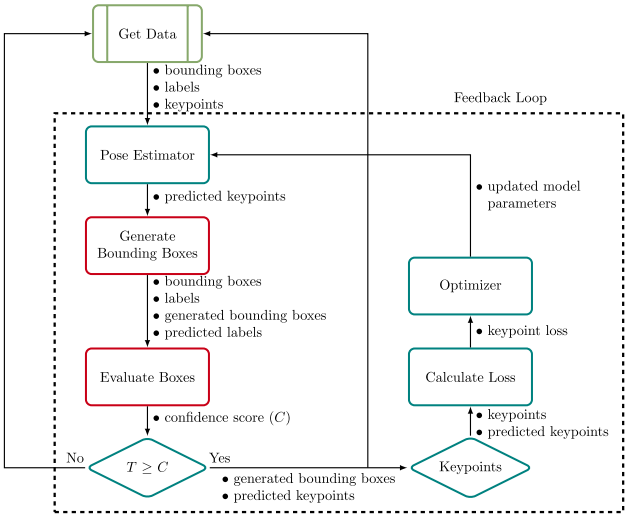


FIGURE 6. KeyEval pipeline feedback loop.

pose estimation, Jaccard Index is referred to as intersection-over-union (IoU). The greater the overlap, the greater the accuracy of the generated bounding box. The mAP metric is calculated by (7), where *AP*, *Precision* and *Recall* are calculated by 8, (9) and (10), respectively. The true positives (TP) and false positives (FP) are determined by the IoU value (11). Using a threshold value, if the IoU of a bounding box is greater to equal to the threshold value, then the bounding box is a TP and if it's less, then it is a FP. The threshold value is a task-specific value, for example, for general object detection, the threshold value typically ranges from 50%-70%. Precision represents the number of correct positive predictions while recall represents the proportion of the positive labels identified. However, for object detection tasks, such as pedestrian detection, the objective is to both detect the object and it's location within the image. Therefore, the mAP metric is commonly used for such tasks as it uses the intersection-over-union (IoU) metric for calculating accuracy. The IoU metric compares the predicted bounding boxes with ground-truth bounding, where the bounding boxes represent the location of the predicted object within the image. The mAP metric is the sum of the average precision (AP) of each class (k) divided by the total number of classes (n). The COCO [19] definition of *AP* is commonly used for evaluating methods that utilise the COCO dataset, such as the DEKR pose estimator that is utilised in this chapter. According to this definition, *AP* is calculated by 8, where the precision values (p) is taken for 11 recall values (r). These recall values range from 0.0 to 1.0, with 11 increments of 0.1. The mAP is calculated by (8), where N is the total number of classes and k is the class.

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$
$$|X \cap Y| = area\ of\ intersection$$
$$|X \cup Y| = area\ of\ union \tag{6}$$
$$mAP = \frac{1}{N}\sum_{k=1}^{N} AP_k \tag{7}$$
$$AP = \frac{1}{11} + \sum_{pr\in\{0,0.1,\cdot,0.9,1\}} p(r) \tag{8}$$
$$Precision = \frac{TP}{TP+FP} \tag{9}$$
$$Recall = \frac{TP}{TP+FN} \tag{10}$$
$$IoU = \frac{area}{union} \tag{11}$$

where,

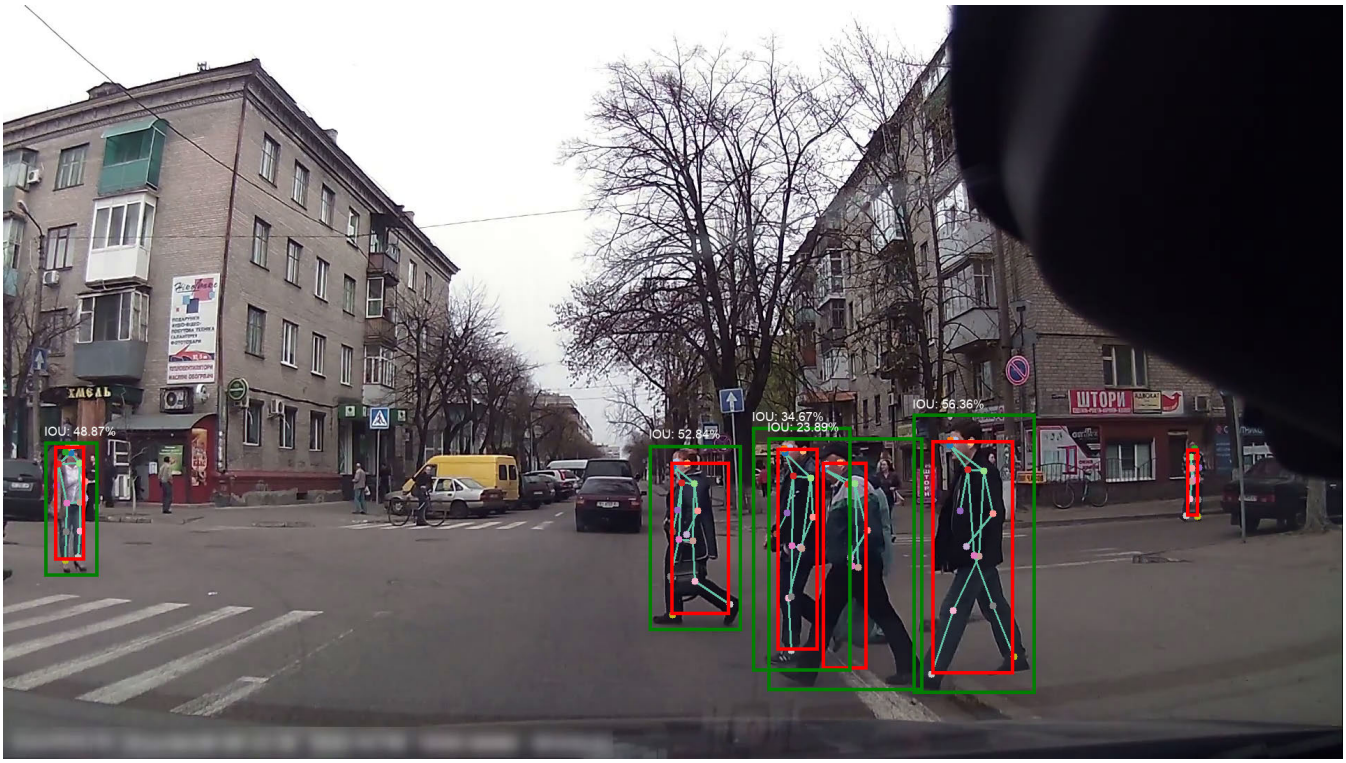
- area is defined by (12)
- union is defined by (13)

$$\begin{aligned} overlap &= (\hat{x}_2 - \hat{x}_1) * (\hat{y}_2 - \hat{y}_1) \\ \hat{x}_1, \hat{y}_1 &= (max(x_1), max(y_1)) \\ \hat{x}_2, \hat{y}_2 &= (min(x_2), min(y_2)) \end{aligned} \tag{12}$$

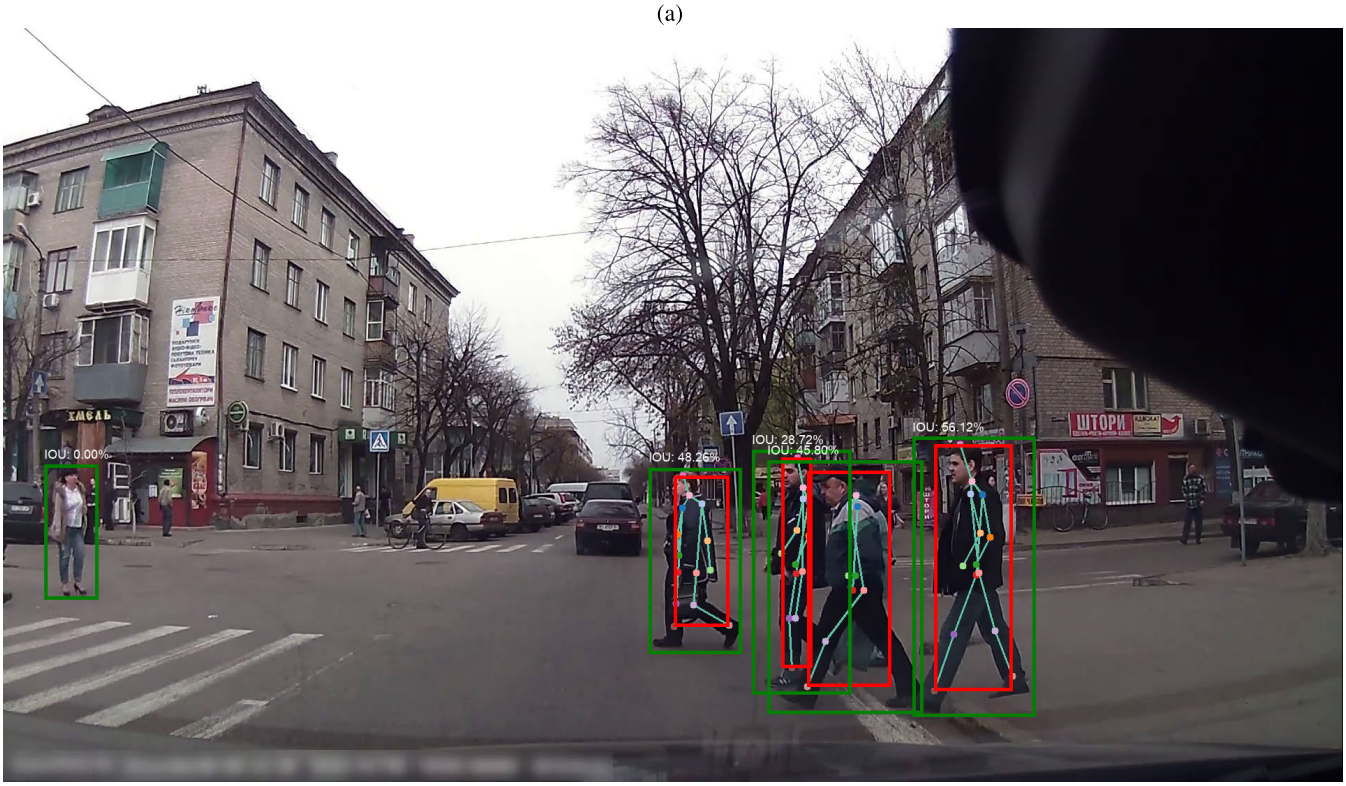
$$union = area(p) + area(g) - overlap$$
$$p = predicted\ and\ g = ground - truth \tag{13}$$

B. EFFECTIVENESS OF FINE-TUNING

Before fine-tuning the DEKR model, a baseline needs to be created. Comparing the performance of the fine-tuned model with the baseline illustrates the improvements made by fine-tuning. The DEKR model is pre-trained using the COCO and CrowdPose datasets and various other settings. One of those settings is the backbone. The backbone is the feature extractor, which is comprised of several channels, which are represented as "w32" and "w48" in Table 2. As the formation of these backbones is out of the scope of this paper, their architecture will not be further discussed in this section. For more information on the formation of these backbones, please refer to [5]. The authors in [5] found that backbones with more channels (i.e., the "w48" variant) are more useful for medium and large pedestrians. This was also true for smaller pedestrians, where the larger number of channels also provided higher resolution of the smaller pedestrian instances. To demonstrate this and also to create the baseline, in Table 2 various backbone settings and their performance for both the JAAD and PIE datasets are compared. The performance is determined by the number of pedestrians detected. The detection is based on the accuracy of generated bounding boxes and ground-truth bounding boxes. The IoU threshold was set to 0.7 and 0.3 for larger and smaller pedestrians, respectively. These values were obtained through a trail-and-error approach, where we balance the total number detections versus the number of correct detections using various threshold values. It was found that predicting keypoints for smaller pedestrians was more difficult, however, using the IoU threshold of



(a)



(b)

FIGURE 7. **Comparing keypoint-based bounding boxes: 7a and 7b illustrates the accuracy of COCO and CrowdPose pre-trained DEKR model, respectively.**

0.3 was sufficient. In this way, only the highly accurate keypoints are utilised for fine-tuning. The default settings for multi-scale pose estimation are implemented for all tests (see Table 1). The DEKR pose estimator is built on top of the

TABLE 2. Baseline for DEKR settings using various settings.

Dataset	Backbone	Weights	Reasonable	Reasonable (small)	All
JAAD	hrnet_w32	COCO	0.2616	0.0887	0.1751
JAAD	hrnet_w48	COCO	0.3101	0.1233	0.2167
JAAD	hrnet_w32	CrowdPose	0.2122	0.0712	0.1417
JAAD	hrnet_w48	CrowdPose	0.2951	0.1211	0.2081
PIE	hrnet_w32	COCO	0.2703	0.1186	0.1945
PIE	hrnet_w48	COCO	0.3318	0.1533	0.2425
PIE	hrnet_w32	CrowdPose	0.2311	0.0781	0.1546
PIE	hrnet_w48	CrowdPose	0.2927	0.1221	0.2074

TABLE 3. Fine-tuning accuracy results for JAAD dataset.

Iteration	Reasonable	Reasonable (small)	All
0	0.3101	0.1233	0.2167
10	0.3324	0.1333	0.2329
20	0.3704	0.2405	0.3055
40	0.4517	0.2809	0.3663
60	0.5343	0.3376	0.4360
80	0.6285	0.3734	0.5001

TABLE 4. Fine-tuning accuracy results for PIE dataset.

Iteration	Reasonable	Reasonable (small)	All
0	0.3318	0.1533	0.2425
10	0.3555	0.1983	0.2769
20	0.3890	0.2571	0.3231
40	0.5134	0.3304	0.4219
60	0.6704	0.3712	0.5208
80	0.7127	0.4385	0.5756

TABLE 5. Keypoint-based bounding box evaluation (AP).

Pose Estimator*	AP^{50}	AP^{30}	AP^{S}	AP^{M}	AP^{L}
JAAD ^{b}	0.533	0.611	0.293	0.612	0.622
JAAD $^{s+f}$	0.640	0.716	0.342	0.689	0.723
PIE ^{b}	0.512	0.569	0.211	0.565	0.593
PIE $^{s+f}$	0.674	0.755	0.376	0.734	0.754

* b - baseline, s - scaled, f - fine-tuned

HigherHRNet [13], as such the backbone is represented by "hrnet_w32" and "hrnet_w48".

Fine-tuning is a technique designed to improve the accuracy of a pre-trained model. Therefore, using the keypoints with the highest accuracy, the DEKR pose estimator can be fine-tuned to predict accurate keypoints for the JAAD and PIE datasets. From the findings in Table 2, the optimal results were for the COCO pre-trained model using the "w48" backbone. Fine-tuning this baseline to improve the accuracy of the predicted keypoints was explored in Table 3 and Table 4. Fine-tuning was performed in iterations of 0, 10, 20, 40, 60, and 80 where an iteration was defined as processing the full dataset and fine-tuning the model. 0 is the baseline, i.e., results without fine-tuning. With each iteration, the results improved, until the performance began to plateau, around iteration 80. The results in Table 3 and Table 4 are of COCO pre-trained DEKR model with the "w48" backbone as it provided the optimal results (see Table 2).

TABLE 6. Keypoint-based bounding box evaluation (AR).

Pose Estimator*	AR^{50}	AR^{30}	AR^{S}	AR^{M}	AR^{L}
JAAD ^{b}	0.266	0.323	0.132	0.342	0.323
JAAD $^{s+f}$	0.332	0.452	0.341	0.533	0.546
PIE ^{b}	0.278	0.299	0.093	0.321	0.332
PIE $^{s+f}$	0.368	0.509	0.378	0.568	0.644

* b - baseline, s - scaled, f - fine-tuned

TABLE 7. Pose estimation evaluation.

Pose Estimator	Reasonable	Reasonable (small)	All
RMPE [20]	0.342	0.012	0.177
HRNet [4]	0.443	0.015	0.229
OmniPose [21]	0.495	0.017	0.256
I^2R-Net [22]	0.503	0.023	0.263
This work	0.736	0.359	0.504

As demonstrated in Table 3 and Table 4, fine-tuning improves the number of pedestrians the DEKR pose estimator accurately predicts. The accuracy went up by 28% for the JAAD baseline (i.e., before fine-tuning) and the best results after fine-tuning. For the PIE dataset, there was a larger accuracy improvement of 33%.

C. GENERATED BOUNDING BOX EVALUATION

Fig. 7a and Fig. 7b illustrate the keypoint-based bounding boxes. The skeleton of the pedestrian is created by connecting all the keypoints. The skeleton is only for illustrative purposes, and unlike in [7], is not required by the proposed approach. The IoU represents the confidence score. The higher the score, the more accurate the generated bounding boxes (blue) when compared to the ground-truth bounding boxes (green). As seen in these examples, the generated bounding boxes are smaller than the ground-truth bounding boxes. For the ground-truth bounding boxes, the full body is considered, while the generated boxes are much tighter to the body and therefore smaller. Fig. 7a and Fig. 7b also demonstrate the accuracy of the COCO and CrowdPose pre-trained variants of the DEKR pose estimator, receptively. As can be seen, the DEKR model pre-trained on the COCO dataset yields more accurate bounding boxes than the model pre-trained on the CrowdPose dataset. One of the reasons could be that the COCO dataset predicts facial keypoints (e.g., ears, nose, eyes), whereas CrowdPose only predicts top-of-head and neck keypoints. This is crucial as the bounding boxes are dependent on the location of the keypoints.

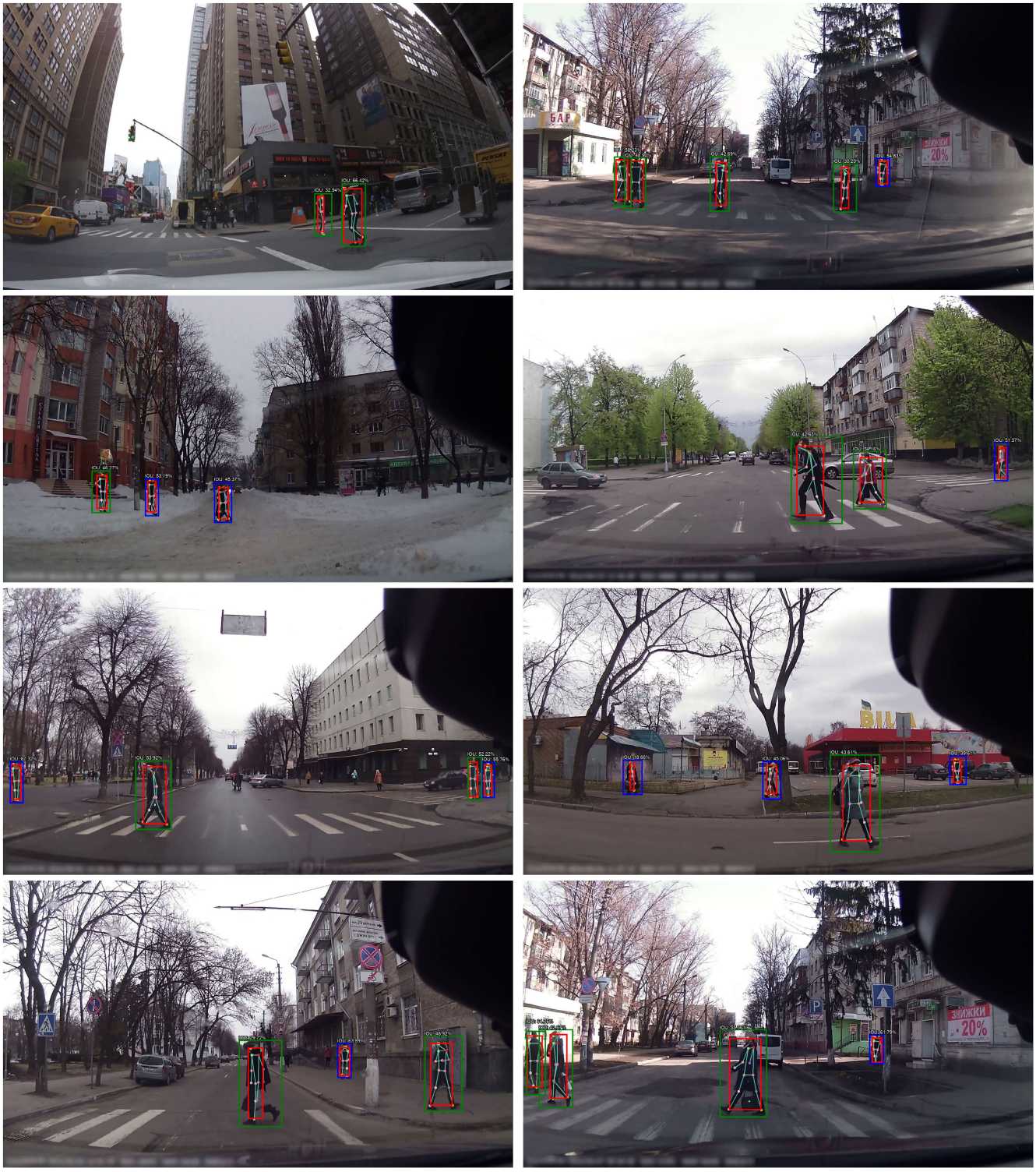


FIGURE 8. KeyEval Prediction Examples: These examples demonstrate the IoU values of detected pedestrians.

As previously mentioned, typically an IoU score of over 50% is considered a good prediction. However, as demonstrated in Fig. 7a and Fig. 7b, the keypoint-based bounding boxes tend to be smaller in area than the ground-truth bounding boxes. Through testing, we found an IoU threshold set to 30% provided the maximum number of correct detections. Higher threshold values omitted many correct detections, and low thresholds did not capture the full pedestrian. Furthermore,

as illustrated in Fig. 7, the keypoint-based bounding boxes tend to be tighter to the pedestrian's siloutette, resulting in smaller bounding boxes. The results are presented in Table 5 and Table 6, where scaled images (based on Section IV-B) and the fine-tuning (see Section V-B) are also considered. These results are calculated using the widely used Average Precision (AP) and Average Recall (AR) metrics. The scaled images are used to improve results for smaller pedestrian instances. As discussed in Section V-B, fine-tuning is designed to improve the overall performance of the proposed approach. The proposed KeyEval architecture achieved accuracy of up to 72% and 76% for the JAAD and PIE datasets, respectively. This is an improvement of 11% and 19% from the baseline for each dataset.

D. COMPARISON WITH STATE-OF-THE-ART

As the JAAD and PIE datasets do not include keypoint annotations, it is difficult to compare the proposed KeyEval method. Therefore, to demonstrate it's effectiveness, the KeyEval pipeline swapped the DEKR pose estimator, with a number of recent pose estimators (see Table 7). Table 7 compares the effectiveness of accurately detecting pedestrians for the *reasonable*, *reasonable (small)* and *all* settings. The results are based on the evaluation technique described in Section IV. The JAAD and PIE datasets were combined for training and evaluation. The JAAD dataset was split with the first 300 videos for training and the last 46 videos for evaluation. The PIE dataset was split with a similar ratio. The training and evaluation samples for each dataset were then combined. Both the aforementioned scaling and fine-tuning techniques were also utilised. As previously discussed, most of the state-of-the-art pose estimators focus on *reasonable* pedestrians, and because of this, the pose estimators in Table 7 struggled to accurately predict keypoints for *reasonable (small)* pedestrians. Therefore, the proposed KeyEval pipeline with the DEKR pose estimator provides optimal performance of multi-scale settings while maintaining comparable performance with current state-of-the-art pose estimators. The KeyEval pipeline outperformed the next best technique in each category. It achieved an improvement of 24% for the *All* setting, where pedestrians of all heights are considered.

E. LIMITATIONS

As only highly accurate keypoints are fed back into the model as ground-truth, when training the model learns these keypoints. This is turn aids the model to predict key-points with higher accuracy after each iteration. This is because each iteration results in correctly predicting more keypoints, which provide further samples which replace the original ground-truth samples. This process continues until the ability of the model to learn further from the samples and the learning begins to plateau (as in Tables 3 and 4). Although, it is demonstrated in this Section that the fine-tuning techniques presented in this paper do indeed improve the accuracy of the pose estimator by nearly 30% when compared to pose estimators which are not fine-tuned, this approach is still limited as not all the pedestrians are accurately detected. Ground-truth keypoints annotations for the JADD and PIE datasets would be ideal for such a use-case as in this paper, where pedestrian intent prediction is investigated using pose information. This would provide keypoints for each pedestrian in the images, regardless of size or shape, particularly for smaller pedestrians, which would provide a larger number of samples for training. This could result in overcoming the plateauing that was previously discussed. The KeyEval approach struggled with smaller pedestrians when compared to larger pedestrians. Although the KeyEval approach presented in this chapter is not optimal, it does provide a means of using datasets lacking keypoint annotations without painstakingly annotating the keypoints by hand, which is a difficult and time-consuming task.

VI. QUALITATIVE ANALYSIS

As demonstrated in Tables 5 and 6, using a threshold value of 30% provides the optimal results. In Fig. 8, a number of examples are provided for multi-scale pedestrians. Some of the examples have an IoU value of around 30%. This typically occurs when a pedestrian is detected while standing straight, reducing the size of the generated box. This reduces the IoU when compared to ground-truth bounding boxes, even though the pedestrian has been corrected detected. Therefore, the value of 30% for the IoU was justified in this scenario.

VII. CONCLUSION

This paper presents the novel KeyEval Pipeline, a approach to generate high-quality human keypoint data. We have demonstrated the effectiveness of keypoint information for predicting crossing intentions of pedestrians in AV applications in our previous work. Our work is intended to be utilise in real-time applications to improve pedestrian safety. Predicting the intentions of pedestrians provides the vehicle with the information required to make adjustments to its path to prevent vehicle-pedestrian incidents.

In this paper, we discuss the KeyEval Pipeline in further detail. Firstly, this approach provides a pipeline for evaluating predicted keypoints for datasets that do not include ground-truth keypoint annotations. Secondly, this pipeline also provides a method of fine-tuning pre-trained pose estimators for improving the quality of predicted keypoints. The KeyEval Pipeline was applied using the state-of-the-art multi-scale DEKR pose estimator. The proposed pipeline provided a novel baseline for multi-scale pedestrian detection on the JAAD and PIE datasets. Using fine-tuning techniques, the baseline was improved upon, achieving an accuracy of up to 76% AP, an improvement of 19% from the baseline. This pipeline is a modular architecture, which means that in the future, a more accurate pose estimator can be utilised for improved performance. Data augmentation was also discussed to overcome

challenges in datasets, particularly the class imbalance issues between crossing/not crossing classes and scale variance of pedestrians.

For future works, we intend to utilise the keypoints generated by the KeyEval Pipeline for gesture recognition (e.g., waving, nodding, etc.) which can further improve our previous techniques that employ keypoint data.

REFERENCES

[1] S. Ahmed, A. A. Bazi, C. Saha, S. Rajbhandari, and M. N. Huda, "Multi-scale pedestrian intent prediction using 3D joint information as spatio-temporal representation," *Expert Syst. Appl.*, vol. 225, Sep. 2023, Art. no. 120077. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417423005791

[2] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 206–213.

[3] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6261–6270.

[4] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696. [Online]. Available: https://github.com/leoxiaobin/

[5] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang, "Bottom-up human pose estimation via disentangled keypoint regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14671–14681.

[6] Z. Fang and A. M. López, "Is the pedestrian going to cross? Answering by 2D pose estimation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1271–1276.

[7] J. Gesnouin, S. Pechberti, G. Bresson, B. Stanciulescu, and F. Moutarde, "Predicting intentions of pedestrians from 2D skeletal pose sequences with a representation-focused multi-branch deep learning network," *Algorithms*, vol. 13, no. 12, pp. 1–23, Dec. 2020.

[8] A. P. Samant, K. Warhade, and K. Gunale, "Pedestrian intent detection using skeleton-based prediction for road safety," in *Proc. 2nd Int. Conf. Adv. Comput., Commun., Embedded Secure Syst. (ACCESS)*, Sep. 2021, pp. 238–242.

[9] S. Ahmed, C. Saha, and M. N. Huda, "Investigation of action recognition for improving pedestrian intent prediction," in *Proc. Annu. Conf. Towards Auto. Robotic Syst.*, 2023, pp. 101–113.

[10] F. Schneemann and P. Heinemann, "Context-based detection of pedestrian crossing intention for autonomous driving in urban environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 2243–2248.

[11] T. Chen, R. Tian, and Z. Ding, "Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3096–3102. [Online]. Available: https://github.com/chen289/Visual-GCN

[12] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "RTMPose: Real-time multi-person pose estimation based on MMPose," 2023, *arXiv:2303.07399*.

[13] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5385–5394.

[14] A. Accv, "Recent advances in human pose estimation and tracking," *Int. J. Comput. Eng. Technol.*, vol. 15, no. 6, pp. 454–463, 2024.

[15] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 264–269.

[16] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 973–986, Apr. 2018.

[17] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[18] L. G. Galvao, M. Abbod, T. Kalganova, V. Palade, and M. N. Huda, "Pedestrian and vehicle detection in autonomous vehicle perception systems—A review," *Sensors*, vol. 21, no. 21, p. 7267, 2021. [Online]. Available: https://www.mdpi.com/1424- 8220/21/21/7267/htm

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

[20] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2353–2362.

[21] B. Artacho and A. Savakis, "UniPose: Unified human pose estimation in single images and videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7033–7042.

[22] Y. Ding, W. Deng, Y. Zheng, P. Liu, M. Wang, X. Cheng, J. Bao, D. Chen, and M. Zeng, "I^2R-Net: Intra- and inter-human relation network for multi-person pose estimation," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 855–862.

SARFRAZ AHMED received the Ph.D. degree in future transport and cities from Coventry University, in 2023, where his doctoral research focused on multi-scale RGB-thermal pedestrian detection and intent prediction for autonomous vehicle applications. He is currently a Research Fellow in AI safety, automation, and applied AI with Coventry University, where he designs validation metrics for evaluating World Foundation Models and contributes to regulatory and industry standards for the safe use of generative AI in autonomy. His broader expertise spans deep learning, multimodal fusion, 3-D perception, synthetic data generation, model optimization, embedded robotics, and AI assurance. His current research interests include AI safety and alignment, trustworthy deployment of large language models, multimodal grounding, LLM-based scenario generation for autonomous systems, and safety-aligned agentic AI architectures for real-world autonomous applications.

M. NAZMUL HUDA is currently a Senior Lecturer (an Associate Professor) in electronic and electrical engineering at the Brunel University of London, an Elected Member of Senate, and the Former Associate Dean (Student Experience). His research focuses on robotics and artificial intelligence (AI) for healthcare and autonomous systems, aiming to deliver intelligent technologies that improve safety, accessibility, and quality of life. His work has resulted in a patent, best paper awards, and publications in leading journals, such as an Expert Systems with Applications. Alongside his research, he served as the Chair for TAROS 2024, a U.K.'s flagship international robotics conference, and contributes to the field through his editorial role for the "Sensors and Robotics" section of the journal *Sensors* and his membership of the EPSRC Peer Review College.

CHITTA SAHA was a Lecturer/an Assistant Professor with the School of Computing, Electronics and Mathematics, Coventry University, from 2012 to 2022. He is currently an Associate Professor with Birmingham City University. Over the last 12 years, he has initiated and conducted high-quality research activities, continuously producing high-quality outputs in international journals and conferences. He is an Active Researcher with an international reputation in the field of solar PV systems, energy harvesting technology, energy storage, and power electronics areas. He is currently supervising Ph.D. students who are working on battery modeling for electric vehicles and solar PV modeling integration with inverter and energy storage for grid application areas.

Dr. Saha was awarded as a fellow of HEA 2015. He was a research associate/fellow on a £2 million multi-universities EPSRC-funded score project at the University of Nottingham, for four and a half years.

MOHAMMED QUDDUS is currently the Chair of Intelligent Transport Systems with the Department of Civil and Environmental Engineering, Imperial College London. He is renowned internationally for his ground-breaking research in transport safety, autonomous and connected transport, big data analytics, and map-matching for intelligent transport systems (ITS). His seminal articles on AI-based map-matching algorithms have been influential and highly cited by researchers worldwide and implemented by the ITS industry, car manufacturers, and National Highways (U.K.). His research projects have been primarily funded by U.K. Engineering and Physical Sciences Research Council (EPSRC), the National Highways, the Department for Transport (U.K.), and European Union (EU). He has an excellent track record in mentoring and supervising post-doctoral researchers and Ph.D. students. So far, he has authored/co-authored over 140 journal articles, 130 conference papers, six book chapters, and 20 technical reports that have accumulated a total of 16 500 citations from researchers all around the world.

Dr. Quddus served as an Associate Editor for a prestigious journal *Transportation Research Part C: Emerging Technologies*.