Unsupervised Pedestrian Pose Prediction
– A deep predictive coding network based approach for autonomous vehicle perception

Xiaoxiao Du[1], Ram Vasudevan[2], and Matthew Johnson-Roberson[1]

***Abstract*— Pedestrian pose prediction is an important topic related closely to robotics and automation. Accurate predictions of human pose and motion can facilitate a more thorough understanding and analysis of human behavior, which benefits real-world applications, such as human-robot interaction, humanoid and bipedal robot design, and safe navigation of mobile robots and autonomous vehicles. This article describes a deep predictive coding network-based approach for unsupervised pedestrian pose prediction from two-dimensional (2D) camera imagery and provides experimental results on two real-world autonomous vehicle data sets. This article also presents a discussion on topics for future work in unsupervised and semi-supervised pedestrian pose prediction and its potential applications in robotics and automation systems.**

I. Introduction

Nowadays, robots exist in an environment filled with moving people. A museum tour-guide robot such as [1], [2] is constantly surrounded by crowds of pedestrians, some needing guidance regarding various exhibitions and some who may intentionally walk up to it or block its way to "test" the system. A self-driving car or a delivery robot can regularly find itself navigating a business district while trying to avoid a collision in the midst of heavy pedestrian traffic. In industrial settings, collaborative robots on assembly lines must allow humans to stand in its proximity and learn to recognize the intention of human workers based on their movements and gestures to ensure safety and improve worker ergonomics and productivity [3], [4]. As these examples suggest, it is important for a robot or automation system to correctly understand and predict human poses to determine the intention of humans and make appropriate decisions and actions to avoid collision as well as facilitate human-robot collaboration.

The human pose is commonly represented using a kinematic skeletal system of body joint locations, or "keypoints," in the perception community. Given sequences of human poses extracted from collected data, such as images or videos in the past, the goal of pose prediction is to capture the underlying motion model of humans and make inferences about the pose and location of a person or persons at future time-steps. There are a number of challenges associated with pedestrian pose estimation and prediction, including body part occlusion by other pedestrians or vehicles, human appearance and clothing variety, data collection range, camera viewpoint and lighting conditions, complex traffic scene in the background, and stochasticity in human motion [5]–[7].

Various machine learning and deep learning methods have been proposed for human pose prediction based on sequences of motion data, including probabilistic dynamic models, physics-based models, supervised deep neural networks, and methods based on video generation. Deep learning (DL) methods such as Recurrent Neural Networks (RNN), in particular Long Short-Term Memory (LSTM) networks [8], have shown superior performance compared with traditional (non-deep) machine learning and physics-based methods for pose prediction. However, supervised DL methods for pose prediction in the literature typically require accurate skeleton annotations of training sequences, which may be difficult or expensive to obtain. If prior annotations are imprecise, erroneous, or contain missing frames, the performance of supervised DL methods may become sub-par. In this work, we developed an unsupervised method for pedestrian pose prediction given car-mounted monocular camera videos collected by an autonomous vehicle. Our proposed system uses a deep predictive coding network (PredNet) [9] as an unsupervised video prediction method to generate future predicted frames and perform image-based pedestrian pose detection, which eliminates the need for accurate measurements and prior skeleton annotations as required by standard supervised pose prediction methods.

II. PredNet-based Unsupervised Pedestrian Pose Prediction

This section describes the foundation of our future frame generation module, PredNet, and our proposed unsupervised pose prediction pipeline.

A. PredNet

The PredNet [9] is a deep neural network inspired by the concept of "predictive coding" from the neuroscience literature. The PredNet architecture consists of four basic layered units: recurrent representation (R), prediction ($\hat{A}$), the input convolutional layer (A), and error representation (E). The right-hand column of Figure 1 shows an illustration of the four-layer PredNet structure used in our system. Given a sequence of images (video data) as input, the target value of the lowest layer A_0 is set to be equal to the image sequence. PredNet goes through a top-down and a bottom-up pass to calculate the states of all units in each layer. First, at the top-most layer ($l = 3$) at timestep t, the recurrent prediction

This work was supported by a grant from Ford Motor Company via the Ford-UM Alliance under award N022884.

[1]X. Du and M. Johnson-Roberson are with Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor, MI 48109 USA xiaodu@umich.edu; mattjr@umich.edu

[2]R. Vasudevan is with the Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109 USA ramv@umich.edu

Digital Object Identifier (DOI): 10.1109/MRA.2020.2976313

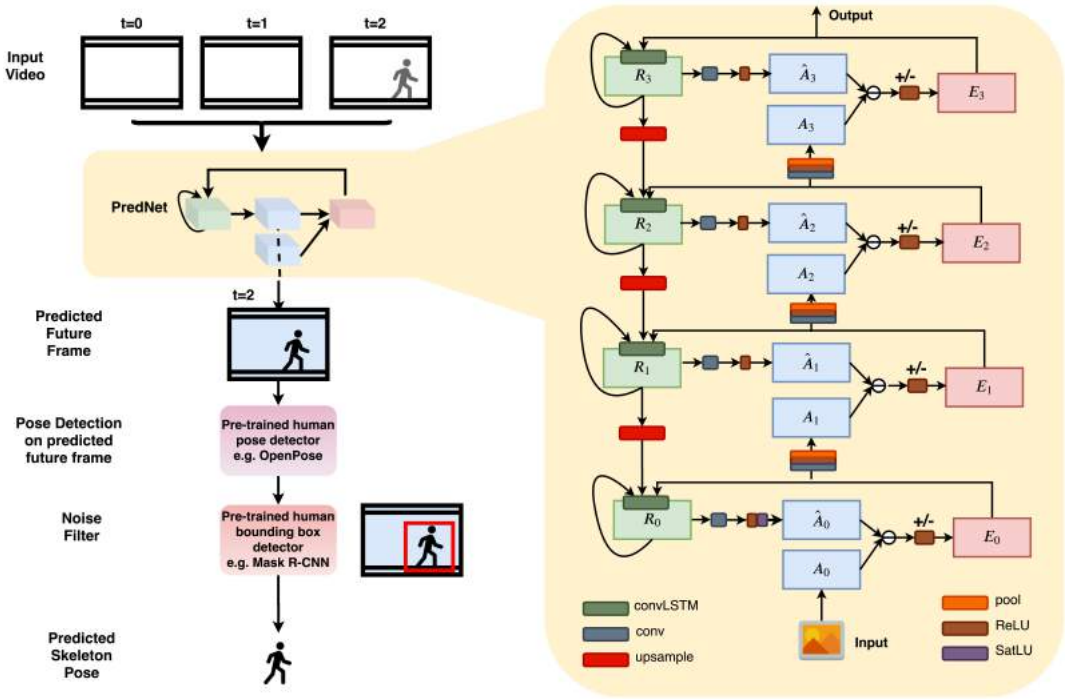

Fig. 1: The system design for the proposed PredNet-based unsupervised pedestrian pose prediction network. The left column shows the flowchart of the proposed system. For this illustration the input video has three timesteps ($t = 0, 1, 2$), but the input video can have more than three frames. The right column shows the details of the four-layer PredNet module where the green, blue and red blocks represent the recurrent representation (R), prediction ($\hat{A}$) and input convolutional layer (A), and error representation (E).

state R_3^t is updated by passing a copy of the error signal at previous timestep E_3^{t-1} and the recurrent prediction state at previous timestep R_3^{t-1} through the convolutional LSTM units. For all subsequent layers $l = 0, 1, 2$, the recurrent prediction state R_l^t is updated according to E_l^{t-1}, R_l^{t-1}, and an up-sampled copy of R_{l+1}^t. At each layer, the prediction $\hat{A}_l^t$ is computed by a convolution of R_l^t followed by a ReLU for non-linearity. The bottom layer $\hat{A}_0^t$ is additionally passed through a saturating linear unit (SatLU) to make sure the predictions of the next frame do not exceed the maximum pixel value. The error response at the bottom layer E_0^t is calculated by passing the difference between $\hat{A}_0^t$ and A_0^t (predicted image and actual image) through a ReLU and then divided into positive and negative errors and concatenated. For $l = 1, 2, 3$, the errors from the layer below E_{l-1}^t are passed through a convolution unit, followed by ReLU and max pooling, to become the input to the next layer A_l^t. The PredNet can be trained end-to-end using gradient descent by minimizing the firing rates of the error neurons. The prediction state of the lowest layer $\hat{A}_0^t$ is returned as the prediction result for timestep t. This process is repeated for all timesteps $t = 1 \rightarrow T$ in the given image sequence. The psuedo-code of the PredNet update process can be seen in Algorithm 1 in [9].

Our pose prediction network is based on, but not limited to, PredNet. In fact, the video prediction module (shaded yellow) can be easily swapped with other video prediction methods. We favor PredNet in our system as it has demonstrated superior performance for future frame prediction with moving backgrounds (as demonstrated in the original paper on the KITTI benchmark suite [10]), which fits our goal of predicting pedestrian poses from data as typically observed from an autonomous vehicle perception system. PredNet is also an unsupervised method, which has the advantage of not requiring any annotation on skeleton or mesh of persons in the scene in prior sequences. In the following experiments, we used a four-layer PredNet structure with 3×3 convolutions and layer channel sizes of (3,48,96,192), and the same parameter settings as described in [9] that produced effective results on natural image sequence prediction.

B. Pose Prediction

Figure 1 provides an illustration for the full system for our proposed PredNet-based unsupervised pedestrian pose prediction network. Given a sequence of RGB images from a camera video as input, the PredNet is used to predict pixel-level RGB values for the next frame. Then, based on predicted future frames, poses were extracted using pre-trained human pose detectors. In our experiments, we used the OpenPose human keypoint detector [11] to detect twenty-five 2D keypoints of pedestrians based on predicted future frames[1]. OpenPose is a deep CNN-based human pose and keypoint detector and has shown state-of-the-art performance in real-time, skeleton-based pose detection based on videos and images containing multiple persons in various actions,

[1]OpenPose for image-based pose detection is available at: https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/quick_start.md#quick-start.

including common pedestrian actions such as standing and walking. To further filter out false alarms caused by non-human objects, such as tree branches, utility poles, and billboards, another layer of human detector is applied to the OpenPose detection results. In our system, we applied a pre-trained Mask R-CNN method [12][2] to estimate bounding box locations for humans and, thus, classify humans with non-human objects. The poses detected within the "human" bounding box are returned as the output of this system as the predicted skeleton pose for the generated future frame.

Our system offers a solution to unsupervised pose prediction inspired by human visual learning. Human drivers are able to easily identify pedestrians and moving objects in the environment based on past observations as the car is in motion. Additionally, human drivers do not necessarily require the exact metric of pedestrians in past sequences to know their pose (for example, walking in a direction that could intersect with the car). Our pose detection system generates future frames based solely on a car-mounted camera and performs direct pose estimation per-frame, which no longer requires accurate measurements and annotations for historic skeleton sequences (which often further requires expensive data collection instruments such as 3D LiDAR[3]) as required by prior work. In addition, the PredNet module inherently incorporates implicit models of scene structures and movements of objects such as buildings, streets, and pedestrians as observed from a moving vehicle. Moreover, previous supervised pose prediction methods often perform poorly with imprecise measurements or occluded body parts in previous frames, yet our method naturally alleviates the missing-data problem since our pose detector is performed per-frame and does not rely on previous skeleton detection results.

III. Experimental Results: JAAD and PedX

We present experimental results of our proposed pose prediction approach on two real-world autonomous vehicle perception data sets, Joint Attention for Autonomous Driving (JAAD) and PedX. The JAAD Dataset [13]–[16] contains 346 video clips collected from one of three different high-resolution monocular cameras in North America and Europe. The camera is positioned inside the car below the rear view mirror. The frame rate is 30 frames per second (FPS). The PedX dataset [17] is a large-scale multi-modal dataset for pedestrians collected at intersections in downtown Ann Arbor, MI, USA in 2017. The dataset provides high-resolution stereo images and LiDAR data with manual 2D ground-truth annotations. The data was captured using two pairs of stereo cameras and four Velodyne LiDAR sensors. Figure 2 shows the data collection set-up of the PedX dataset. We only use data from one of the forward-facing cameras with 6FPS frame rate in the following experiments. Both JAAD and PedX provide car-mount RGB videos in urban traffic scenes and observe pedestrian movements in-the-wild. The JAAD dataset was collected on a moving car (thus, moving background and moving pedestrians), while the PedX dataset was collected on a parked autonomous vehicle (stationary scene, moving pedestrians). The JAAD dataset also includes a variety of traffic scenes such as roads and parking lot, while PedX focuses on three urban intersections.

Fig. 2: The PedX data collection system on an autonomous vehicle, parked at an intersection in downtown Ann Arbor, MI, USA.

The JAAD dataset contains 81,906 frames in total. We divided the JAAD dataset in 30-frame sequences (one second-long) and used randomly-shuffled 85% of the sequences for training, 5% for validation, and tested on the remaining 10% sequences (8,280 frames). We used the same parameters of the PredNet model as described in [9] and trained the model from scratch. The input images were down-sampled to 256×456 to accommodate the computer memory while maintaining the aspect ratio. We also divided the PedX dataset into 30-frame sequences (approximately five seconds long) and used the PredNet module trained from JAAD to test on all 484 PedX sequences (14,520 frames).

A. Frame Prediction Results

We first report the quantitative evaluation of the predicted future frames as compared to the actual collected video frames. Two metrics are used, mean-squared-error (MSE) and the Structural Similarity Index Measure (SSIM) [18]. The MSE calculates the pixel difference between the predicted and actual frames (the lower the better) and the SSIM is correlated with perceptual similarity (the larger the better). We also report the results from the trivial solution of copying the last frame as baseline. As shown in Table I, the PredNet module is effective in predicting the future frame and achieves better performance in both MSE and SSIM in both JAAD and PedX data sets. The MSE is lower and the SSIM is higher for the PedX dataset overall since the background scene in PedX data is fixed while the background changes as the car moves in the JAAD dataset.

Figure 4 provides visual examples of the future frame prediction results of PredNet on sample sequences from JAAD and PedX. We observed that the frames predicted by PredNet (row 2 and row 5) were very similar visually to frames from the actual video clips (row 1 and row 4). The frame prediction step was also capable of extrapolating

[2]Pre-trained COCO weights for Mask R-CNN human detection are available at https://github.com/Superlee506/Mask$_RCNN_Humanpose/releases$.

[3]Light Detection and Ranging, an instrument commonly used for autonomous vehicle perception to measure precise ranges of objects. High-resolution LiDARs can be expensive.

sky and textures of other objects and making fairly accurate predictions, such as when cars move in or out of the frame.

TABLE I: Evaluation of next-frame predictions on JAAD and PedX data sets. The best performance is in bold and the standard deviation across three runs is presented in parentheses in the following tables.

	MSE ($\times 10^{-3}$)	SSIM
PredNet on JAAD	**2.621(0.587)**	**0.926(0.017)**
Copy Last Frame on JAAD	6.602(0.732)	0.904(0.028)
PredNet on PedX	**1.244(0.093)**	**0.962(0.003)**
Copy Last Frame on PedX	1.759	0.955

B. Pose Prediction Results

The output of PredNet frame prediction is of shape $(N, T, 256, 456, 3)$, where N is the number of sequences and T is the sequence length (T=30). To ensure optimal performance of the OpenPose keypoint detector, we used the *pyplot.savefig* function in *matplotlib* in Python to save the frame predictions with dpi (dots per inch) equals to 1000, resulting in 2791×4967 images for the JAAD dataset and 2687×3645 for the PedX dataset for pose prediction. We call this process "up-sampling". The OpenPose was then applied to the up-sampled (saved high-resolution) images to detect pedestrian keypoints.

Since the JAAD dataset did not provide skeleton annotations, we used OpenPose pose detection results filtered by Mask R-CNN, based on the actual video frames, as the ground-truth. We then compared the pose detection results from our proposed approach based on the predicted future frames to the ground-truth and calculated the root-mean-square-error (RMSE) results. In PedX, we also report the RMSE results between our pose prediction results and the manual 2D annotations. We compared our approach with two previous supervised deep learning pose prediction methods, LSTM-3LR [6] and a frame difference method adapted from Bio-LSTM (Bio-LSTM-L_c) [19]. For LSTM-3LR, we used a stacked three-layer LSTM with 64 units and 250 training epochs in our experiments. For the frame difference (Frame Diff.) method, we used the most prominent gait feature, the gait periodicity loss (L_c) in the Bio-LSTM objective function, with a two-layer stacked LSTM recurrent neural network for pose prediction. We also reported the RMSE results from the naive baseline by copying poses from the last frame.

Table II presents the pose prediction results on the JAAD dataset. The RMSE-x, RMSE-y, and RMSE-xy columns corresponds to the RMSE results on the x-axis (width of the image), y-axis (height of the image), and average of both, respectively. The lower the RMSE results, the better the prediction performance. The second row shows the RMSE results in pixels. The last row below the double line shows the percentage pose prediction error normalized by bounding box (bb) sizes, and their corresponding error values in metric space in centimeters. The symbol $\hat{=}$ means "corresponds to". As shown, our proposed approach yields smaller RMSE mean and standard deviation values than comparison methods in both x and y (width and height) directions,

TABLE II: Evaluation of pose prediction on the JAAD dataset.

	RMSE-x	RMSE-y	RMSE-xy
Copy Last Frame	102.534(165.329)	89.801(137.015)	101.561(148.418)
Frame Diff.	92.456(151.035)	78.352(117.988)	89.816(132.828)
LSTM-3LR	72.366(125.301)	57.324(95.502)	68.374(109.531)
Ours	**29.347(39.818)**	**19.088(21.685)**	**26.516(30.623)**
Ours: bb	0.145(0.144) $\hat{=}$ 14.5(14.4)cm	0.040(0.042) $\hat{=}$ 8.0(8.2)cm	-

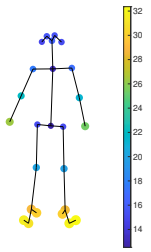

Fig. 3: The predicted skeleton RMSE results of our proposed method, in pixels. Each circle represents a keypoint (a joint) on the body. The 25 keypoints include left and right shoulders, elbows, wrists, hips, knees, ankles, eyes, ears, feet, toes, heels; and nose, neck, and center of hip. The colors of the circle represents the mean RMSE results across three runs; yellow shows higher error and dark blue shows low error. The size of the circle represents the standard deviation value at each joint; the larger the circle, the higher the standard deviation.

indicating the effectiveness and consistency of the proposed method. The two supervised comparison methods, LSTM-3LR and Bio-LSTM, were both originally developed for 3D skeleton/mesh prediction and require accurate full-body annotations in previous frames. In this experiment, where 2D keypoints detected in previous frames are often imprecise or missing due to occlusion, those supervised methods produced higher error. The trivial solution of directly copying the last frame yielded the highest error in both height and width directions, which is as expected since this baseline method does not account for pedestrian motion in between frames at all. Our proposed pose prediction method, on the other hand, produces pose estimations based on the predicted future frame and does not require previous skeleton detection results, which resulted in significantly smaller error and more consistent pose prediction results.

Figure 3 plots the skeleton RMSE results in pixels at each joint. As shown, the body center has the lowest error whereas the extremities (hands and feet/heels) show both larger mean and standard deviation error values, which is understandable as the hands and feet are the most flexible part of the human body and also most easily occluded, therefore the most difficult to predict. To further "translate" the image-level evaluation (in pixels) to metric space (in meters), we computed the pose prediction errors normalized by the heights and widths of the corresponding bounding boxes detected by Mask R-CNN. The last row in Table II shows the percentage error with respect to human bounding box sizes. Assuming that a person is approximately two meters tall and one meter wide, our average pose error in the physical space is approximately 14.5cm in lateral direction and 8cm height-wise.

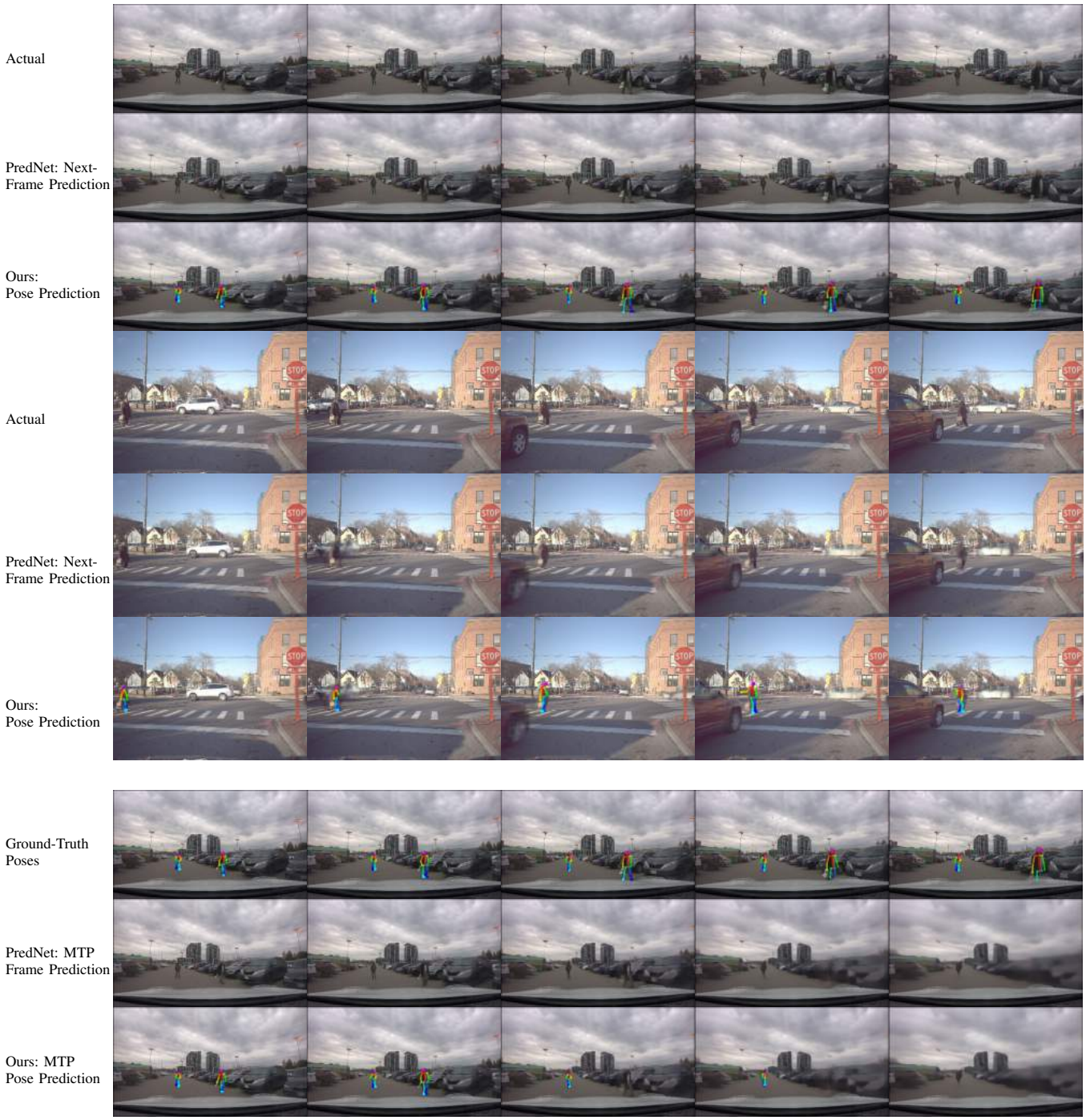

Fig. 4: An example of sample frames from the JAAD and PedX video sequences, the PredNet future frame prediction results, and the pose prediction results of our proposed system. The top six rows show next-frame prediction results. The first three rows are from the JAAD dataset, where two pedestrians are walking in a parking lot, and the middle three rows are from the PedX dataset, where a pedestrian is walking across the crosswalk at an intersection. The bottom three rows show the ground-truth poses from the JAAD dataset, the multiple-timestep frame prediction results, and our MTP pose prediction results. The original video sequences are 30-frames-long, which is equal to one second in JAAD and five seconds in PedX. Here we plot $t = 1, 7, 13, 19, 25$, which corresponds to approximately 0.2-second-spacing in JAAD sequence and 1-second-spacing in PedX. For more results containing consecutive frames, please see supplementary video.

We used the PredNet module trained from JAAD to test on the PedX dataset. Table III presents the pose prediction results on PedX. Since the PedX dataset contains manual ground-truth (GT) annotations, we compared our pose prediction results to both the manual GT and the poses detected by OpenPose on the actual video clip. We also evaluate the accuracy of the OpenPose keypoint detector compared with the manual GT. Compared with Table II, the PedX prediction results have slightly higher RMSE than when tested on JAAD dataset due to the frame rate being five times higher in JAAD than PedX, and that the PedX scenes have never been observed in training. Nevertheless, using the model trained on JAAD can still make pose prediction on previously unseen PedX contexts with relatively small error (25.1cm in width and 12.6cm in height).

Figure 4 show visual examples of pose prediction results for sample sequences from JAAD and PedX. The predicted skeleton poses were overlaid with the predicted RGB images in row 3 and row 6. Our pose prediction system can correctly detect and predict human poses in the scene based on the predicted future frame from PredNet. Particularly, when a pedestrian was partially occluded because of the motion of the ego vehicle (the vehicle with the data collection camera), as shown in the last column of row 3, our pose prediction process was still able to detect the correct pose and mark that the legs are occluded (not plotted on the image).

TABLE III: Evaluation of pose prediction on the PedX dataset. The "OP Actual vs. GT" row shows the RMSE results between pose detected by OpenPose on the actual video clip and the manual ground-truth annotation (GT). The "Ours vs. OP Actual" row shows the RMSE results between our proposed method and poses detected by OpenPose on the actual video clip. The "Ours vs. GT" row shows the RMSE results between the OpenPose pose on the predicted video clip (i.e., our proposed method) and manual GT.

	RMSE-x	RMSE-y	RMSE-xy
OP Actual vs. GT	39.616(49.049)	31.014(45.166)	38.634(52.428)
Ours vs. OP Actual	58.235(33.514)	33.486(22.751)	49.689(24.653)
Ours vs. GT	68.781(47.798)	43.546(40.845)	62.000(45.836)
Ours: bb	0.251(0.148) $\hat{=}$ 25.1(14.8)cm	0.063(0.055) $\hat{=}$ 12.6(11.0)cm	-

C. Multiple-Timestep Prediction

This section presents experimental results of predicting multiple timesteps (MTP) in the future. Given ten actual observed frames in the JAAD dataset, our PredNet-based video generation module extrapolates the next 20 timesteps, in which the 11^{th} frame prediction was fed back to the network as input to generate the 12^{th} frame prediction, and so on. Then, our pose prediction module performs pose estimation on the predicted frames. Since R_l^0 and E_l^0 in PredNet were initialized to zero, the prediction at the initial timestep was spatially uniform and therefore not considered in our analysis.

Figure 6 shows the multiple-timestep frame prediction results evaluated by the aforementioned two metrics, MSE and SSIM. Recall that the MSE calculates the pixel difference between the predicted and actual frames (the lower the better) and the SSIM is correlated with perceptual similarity (the larger the better). The second timestep produced high frame error due to the fact that there is no motion information available yet in the sequence, causing the frame reconstruction to be blurry. In the next few timesteps, PredNet learned the underlying dynamics in the motion sequence and the predicted frames better matched the input (actual) frames, resulting in relatively low MSE and high SSIM results between timesteps 3-10. This observation is consistent with the description in [9]. After time step 10 (in the MTP process), since previous predictions were used as input to extrapolate future frames, the frame errors were higher than next-frame prediction results and the MSE increased (and SSIM decreased) over time.

Figure 7 shows the pose prediction results for time steps 2-30 based on the MTP frame predictions, with comparison to the next-frame prediction results. In time steps 2-10, where the actual frames were used as input, the MTP process is the same as the next-frame prediction and they yield comparable RMSE results. In time steps 11-30, where the previous prediction results were recursively iterated as inputs to generate future frames, the MTP error are higher than next-frame prediction and increased significantly over time as the noise overcame the system. Qualitatively, the bottom three rows in Figure 4 shows the MTP frame and pose prediction results of a sample sequence from the JAAD dataset. As time increases, the MTP model was still able to capture certain motions in the scene, such as the movements of the clouds and surrounding cars in the parking lot. However, the frame predictions eventually became more and more blurry due to the accumulation of noise and uncertainty, thus causing inaccurate and missed pose detection results (such as the person on the right-hand side) in its extrapolations after the first ten time steps.

D. Discussion

The quantitative and qualitative results presented above show that our proposed PredNet-based unsupervised pose prediction approach can produce accurate pose prediction results. Using two real-world data sets in autonomous driving contexts, JAAD and PedX, we show that our proposed approach is applicable to a variety of driving scenes and the frame prediction models are can be generalized to previously unseen environments.

Our PredNet-based frame prediction step produces realistic future frames, accounting for both vehicle and pedestrian motions. We also observed that parts of the predicted frames can be blurry later in the sequences (when t increases) due to down-sampling, particularly in small textured regions, such as the pedestrian's facial features, or unfamiliar regions, such as when the red car entered the scene in the middle of the sequence in Figure 4. In PedX, the blur effect seemed more prominent due to the fact that we used a PredNet model pre-trained on JAAD and the PedX environments were not learnt in training and that the spacing between frames was longer in PedX due to a low frame rate. However, such blur effect did not have a significant impact on the pose prediction step and we observed that our pose prediction module can still

Fig. 5: Computation time analysis for the JAAD dataset experiments, trained and validated on 73,626 frames, tested on 8,280 frames.

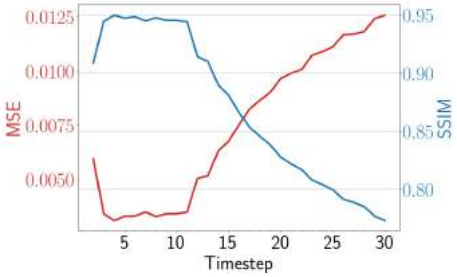

Fig. 6: Multiple-timestep frame prediction results on the JAAD dataset. The x-axis marks the timesteps (the last 20 timesteps were extrapolated from the MTP process) and the y axes show the frame MSE (left, red) and SSIM (right, blue).

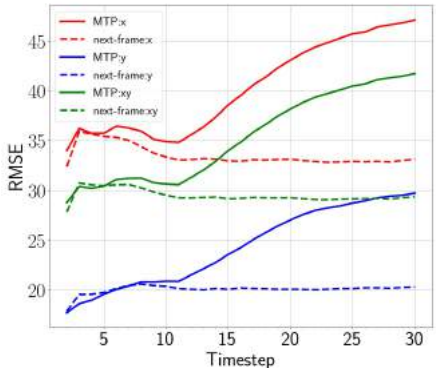

Fig. 7: Multiple-timestep pose prediction results on the JAAD dataset. The x-axis marks the timesteps (the last 20 timesteps were extrapolated from the MTP process) and the y-axis marks the openpose skeleton joint RMSE in pixels. The solid lines show the MTP results and the dashed lines show the next-frame prediction results. The red, blue, and green colors represent the RMSE results on the x-axis (width of the image), y-axis (height of the image), and average of both, respectively.

successfully detect the skeleton joint locations of pedestrians based on predicted future frames.

Our proposed approach does not rely on manually-defined models and distributions but instead performs frame-based 2D pose prediction. Our approach transforms the standard supervised learning problem, which requires full-body pose annotations in prior sequences, into an image-data-driven, unsupervised framework. Our frame prediction step realistically predicts scene dynamics as well as the relative motion between pedestrians (or moving objects) and background, and our pose detection step naturally handles occlusion based on the OpenPose detector. Moreover, our approach produces both the RGB future frame as well as the future pose, which makes the pose visually interpretable within the frame.

One of the challenges associated with such frame-based pose prediction methods is the performance of the pose prediction depends significantly on the accuracy of the human detector, i.e., the OpenPose and Mask R-CNN algorithms used in our system. Although highly effective most of the time, the human detectors can occasionally produce false alarms, such as mistaking a windshield wiper or a billboard painting as a pedestrian, or mis-identifying tree branches or dark shadows as a person, especially under difficult lighting conditions such as sun glare, as shown in Figure 8. It is also challenging at times to accurately identify multiple persons moving in a crowd at a distance. This can be solved by better positioning the camera, such as outside the windshield wiper or on top of the car to avoid potential interference of the wipers, using more vigorous noise filtering approaches, and exploring alternative sensor modalities to compensate the lighting and other environmental effects on RGB cameras.

Fig. 8: Examples of OpenPose false alarms, where a windshield wiper and a person on the billboard were mis-identified as pedestrians (top) and tree branches were mis-identified as pedestrians due to dark shadows under glare (bottom).

E. Analysis on Computation Time

Our proposed network was implemented in Python 3.6 using the Keras framework [20]. The PredNet module was trained on a desktop computer with Intel Xeon 2.10GHz CPU with four NVIDIA TITAN X GPUs and 128 GB memory. The memory requirement can be reduced if the input video is of lower resolution or if the sequence length is smaller. After PredNet was trained, the pose detection module was performed on a desktop computer with Intel i7 3.60GHz CPU with two NVIDIA TITAN X GPUs.

Figure 5 shows a breakdown of our end-to-end computation time for the JAAD dataset experiments (next-frame prediction). The PredNet training and upsampling are the two most time-consuming tasks. However, the PredNet training step can be omitted when a pre-trained PredNet model is applied to other data sets, such as in our PedX experiments. The training time also varies if the input data size changes. The inference time for the PredNet testing stage is approximately 20ms per frame. The OpenPose pose detection step is also quite fast, taking approximately 418ms per frame. In our current implementation, since our pose detector and noise filter are both image-based, we saved all up-sampled

RGB images to disk. Therefore, the upsampling and Mask R-CNN filter steps include the long disk read and write time. Future work will include investigating non-image-based pose prediction methods. Additionally, the pose detection and filtering steps can be trivially parallelized.

IV. Future of Pose Prediction and Conclusions

This article presented an unsupervised pedestrian pose prediction system based on a deep predictive coding network for autonomous vehicle perception. Our system combines video generation approaches, such as PredNet, and frame-based pose detectors, such as OpenPose and Mask R-CNN, and shows effective future pose detection performance on real-world autonomous vehicle applications.

There are several new research directions related to this work. This article focused on autonomous driving perception applications, but human pose prediction can be widely applied in various robotics and automation applications, including virtual reality, sports and artist posture analysis, and medical assistance. In many robotics and automation applications with abundant perception data, accurate annotations are expensive and difficult to obtain. Therefore, it is necessary to further develop unsupervised and semi-supervised learning methods given sparse and imprecise labels for pedestrian pose prediction in various contexts. Particularly, with human pose analysis and sequence prediction, additional investigation can be conducted to incorporate more realistic spatial, temporal, textural, semantic, and biology-derived constraints in the learning model [21]. Furthermore, mobile robots and automation systems are interacting with the real world. In autonomous driving, pedestrians and vehicles are constantly making decisions based on their interactions and finding the balance between achieving certain goals and avoiding risk and collision. Interesting future work will include incorporating pedestrian-pedestrian and pedestrian-vehicle interactions in pose and trajectory prediction and use such prediction results for activity inference, low-level decision making (such as stop/go), and path planning.

Our current pose prediction performance depends on the accuracy of the frame prediction in RGB image space, which may become more challenging as weather and lighting conditions change. Future work will include investigating human pose prediction based on alternative or complementary sensors, such as depth cameras and thermal imaging cameras. With the development of advanced computing units, network architecture design and automated parameter settings for improving the efficiency and scalability of deep learning approaches can be further studied as well.

Finally, it is essential to develop evaluation metrics for pose prediction, particularly when ground-truth is not available or imprecise. An open problem remains as to how to evaluate if the predicted pedestrian poses are natural and realistic and if they are in accordance with the traffic rules and regulations and ethics.

References

[1] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, "The interactive museum tour-guide robot," in *Proc. AAAI Conf. Artificial Intell.*, 1998, pp. 11–18.

[2] S. Thrun, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte *et al.*, "Minerva: A second-generation museum tour-guide robot," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, vol. 3, 1999.

[3] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, 2018.

[4] W. Kim, M. Lorenzini, P. Balatti, D. H. P. Nguyen, U. Pattacini, V. Tikhanoff, L. Peternel, C. Fantacci, L. Natale, G. Metta, and A. Ajoudani, "Adaptable workstations for human-robot collaboration: A reconfigurable framework for improving worker ergonomics and productivity," *IEEE Robot. Autom. Mag.*, 2019.

[5] I. Petrov, V. Shakhuro, and A. Konushin, "Deep probabilistic human pose estimation," *IET Comput. Vis.*, vol. 12, no. 5, pp. 578–585, 2018.

[6] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 4346–4354.

[7] J. Bütepage, H. Kjellström, and D. Kragic, "Anticipating many futures: Online human motion prediction and generation for human-robot interaction," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 1–9.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," in *International Conference on Learning Representations*, 2017.

[10] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[11] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2961–2969.

[13] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint attention in autonomous driving (JAAD)," *arXiv preprint arXiv:1609.04741*, 2016.

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

[15] ——, "Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 206–213.

[16] ——, "Understanding pedestrian behavior in complex traffic scenes," *IEEE Trans. Intell. Veh.*, vol. 3, no. 1, pp. 61–70, 2017.

[17] W. Kim, M. Srinivasan Ramanagopal, C. Barto, K. Rosaen, M.-Y. Yu, N. Goumas, R. Vasudevan, and M. Johnson-Roberson, "Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1940–1947, April 2019. [Online]. Available: http://pedx.io/

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[19] X. Du, R. Vasudevan, and M. Johnson-Roberson, "Bio-LSTM: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1501–1508, April 2019.

[20] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[21] F. Bonsignorio, J. Hallam, and E. M. A. P. del Pobil, "The road ahead-final remarks," *Metrics of Sensory Motor Coordination and Integration in Robots and Animals: How to Measure the Success of Bioinspired Solutions with Respect to Their Natural Models, and Against More 'Artificial' Solutions?*, vol. 36, pp. 181–185, 2019.