International Journal on Advanced Science Engineering Information Technology

Vol.13 (2023) No. 4
ISSN: 2088-5334

# Robust Pose Estimation of Pedestrians with a Deep Neural Networks

Chuho Yi [a], Jungwon Cho [b,*]

[a] *Department of AI Convergence, Hanyang Women's University, Seoul, Republic of Korea*
[b] *Department of Computer Education, Jeju National University, Jeju, Republic of Korea*
*Corresponding author: *jwcho@jejunu.ac.kr*

***Abstract*— In this paper, we provide a method for robust estimation of pedestrian pose that is especially useful for autonomous vehicles traveling toward pedestrians far away. Pedestrians in the far distance appear relatively small when seen by a camera, making it difficult to estimate the pedestrian's pose. We use fused deep neural networks (DNNs) to resolve the problems presented by pedestrians in the far distance. First, DNNs are used to detect pedestrians and enlarge the observed image. Next, the DNN method of pose estimation is applied. The proposed method uses a single camera to estimate the posture of a pedestrian in the far distance. Far-off pedestrians observed by cameras in moving cars appear as low-resolution images of non-rigid bodies. Detection and orientation estimation are difficult with conventional image processing methods. We used a series of DNNs to detect pedestrians, improve data availability, and estimate challenging postures to address these limitations. In this paper, we propose a method based on the multi-stage fusion of DNNs to solve a difficult problem for a single DNN. The experimental results established the superiority of the proposed method when applied to data challenging for conventional pose estimation methods. Applications of the proposed method include observing small objects and objects in the far distance. The method may be especially useful in surveillance systems, sports broadcasting, and other applications requiring human posture estimation.**

***Keywords*— Pose estimation; pedestrian; deep neural network (DNN); super resolution; mono camera-based estimation.**

*Manuscript received 11 Dec. 2022; revised 23 Mar. 2023; accepted 12 May 2023. Date of publication 31 Aug. 2023.*


## I. INTRODUCTION

Driver assistance systems (DAS) are an important safety feature in autonomous vehicles. They can predict pedestrian movement and detect pedestrians on the road, thus helping to prevent accidents. However, when the distance between the car and the pedestrian is relatively large, the small size of the pedestrian in the camera image makes detection challenging. As cars approach far-off pedestrians relatively quickly, detecting these pedestrians and predicting their behavior is a difficult problem. The direction of pedestrian movement is important for predicting pedestrian behavior, as is the direction of their gaze and level of awareness of the oncoming car. Techniques for predicting pedestrian body pose and determining head direction generally require an image of a minimum size. This paper proposes a robust pedestrian pose estimation method using a single camera observing a pedestrian in the distance.

Human pose estimation technologies are used to detect a person's posture in an image or video. Existing technologies determine posture by identifying and connecting key points (the joints of the body). Applications of these technologies include autonomous driving, security, sports, games, augmented reality, and other fields that require recognition of user movements. Pose estimation research focuses on shifting from the top-down to the bottom-up method. The top-down method begins with detecting an object within an image (i.e., identifying and localizing a person in a bounding box). Detection accuracy is higher than that of the bottom-up method because object detection is a distinct step. However, detection speed is relatively low. With the bottom-up method, the first step is to detect and display the keypoints of major body parts; the initial object detection step is omitted. This increases detection speed in comparison with the top-down method but reduces accuracy.

It is time-consuming to apply human pose estimation to an entire image. To solve this problem, Papandrou et al. [1] and Wang et al. [2] proposed a network that carries out two-step pose estimation. Our method used the two-step approach and applied Fast R-CNN to locate a person before cropping and proceeding with human pose estimation. Cao proposed a method using affinity fields and confidence maps [3-6]. First, the image is passed through 10 layers of the vgg-19 network, and a feature map is computed. Next, affinity fields and

confidence maps are generated by using the feature map as an input for stages 1–6. At each stage, loss is calculated by comparing the output with ground truth (GT) labels. The network is trained to optimize the results of these comparisons. After these steps, the confidence maps and affinity fields can be combined to create a complete human skeleton. Data are combined using a greedy relaxation algorithm. Fang proposed three methods for estimating the poses of multiple people in real-time [7-12]. First, a symmetric spatial transformer network (SSTN) optimizes the network using parallel single-person pose estimator (SPPE) branching for high-quality human pose estimation but with inaccurate pedestrian detection. Second, the Parametric Pose Non-Maximum Suppression (NMS) algorithm uses database methods to optimize postural parameters by comparing and removing poses. Third, the Pose-Guided Proposals Generator (PGPG) enhances training data. The human body region can be recreated by learning how various poses are classified based on the model's output.

Generally, human pose estimation aims to provide quick and accurate estimates. The topic of our paper is pose estimation at long distances, which has not been covered in previous studies. To this end, we provide a method for detecting the posture of a pedestrian in a far distance. In these cases, estimating the posture using small images is necessary. A method that can enlarge the image accurately is required; this paper applies the super-resolution (SR) imaging method.

Image SR refers to converting a low-resolution image into a high-resolution image. Image SR can be divided into single-image super-resolution (SISR) and multi-image super-resolution (MISR), depending on whether one or multiple images are used. SR is used to reconstruct high-resolution images from low-resolution images, but multiple 'correct' reconstructions are possible. The lack of a correct answer means the problem cannot be defined. These problems are known as regular inverse problems or ill-posed problems.

To overcome these difficulties, we first define a GT high-resolution target image. Blurring, down-sampling, and noise injection are used to convert this into a low-resolution image. Then, the model is trained to restore the low-resolution image to GT using the seedling method. The performance of SR may vary depending on the distortion and down-sampling techniques used to create the low-resolution image. This reflects the fundamental limitations of SISR.

Ledig conducted a study using a generative adversarial network (GAN) to generate virtual data for SR using arbitrary random numbers between generators and discriminators [13-20]. This research also used a super-resolution GAN (SRGAN), whereas the conventional method uses mean squared reconstruction error to obtain the peak signal-to-noise ratio (PSNR). Although the conventional method is valuable, Ledig notes that it produces a slightly blurred output and suggests a way of using a GAN to restore an image that appears plausible to the human eye. In GAN-based methods like SRGAN, distortion indicates low performance; however, user satisfaction is improved compared to existing SR methods.

In this paper, we use three steps to estimate the posture of a pedestrian in the far distance. First, edge and lane information and other observations are used to obtain the vanishing point, define the far-distance area, and expand the data. I the object is estimated to be a long distance away, part of the image is enlarged. The SRGAN is used instead of simple up-sampling. Next, the CNN method is used to generate feature maps. The feature map detects the key points of major body parts of the pedestrian and estimates the pedestrian's pose. The effectiveness of our method is confirmed through the experiments described in this paper.

## II. Materials and Method

The method proposed in this paper is depicted in Figure 1. In module (a), vanishing point estimation is performed, and a far-distance area of an image perceived by an autonomous vehicle is selected. A region of interest is defined because the subsequent procedure uses relatively intensive calculations. In module (b), the pedestrian is identified using Fast R-CNN. GAN is used in module (c), and SR is used to enlarge small images in far-distance areas. Finally, in module (d), confidence maps, feature maps, and affinity fields are calculated to obtain the final pose estimation.

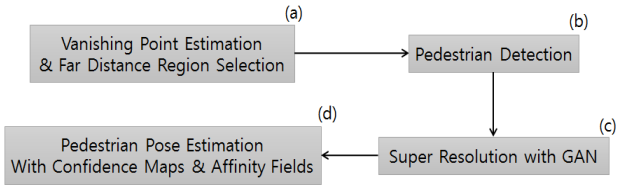



Fig. 1 The proposed method.

### A. Far-Distance Region Selection with DNNs

Estimates of vanishing points from the front camera of an automobile can vary widely depending on the car's pitch, the condition of the road surface, and the shape of the road ahead. Information on lane features is very important and is obtained through edge information from surrounding areas. We use Lee's method [21]-[24] in this paper. Lee notes that there is a large difference in illumination between the daytime and nighttime and that changes in the weather and in-vehicle operation (e.g., the use of windshield wipers) make the vanishing point difficult to measure through video processing. Using a vanishing point guided network (VPGNet), we can robustly determine far-distance regions, the vanishing point, and lane information under various road and weather conditions. Selection of the vanishing is difficult when it is based on a single image taken with a mono-camera mounted on a car, due to vibration. We used the Kalman filter to solve the vibration problem and select the far-distance region [25], [26]. Figure 2 shows the result.

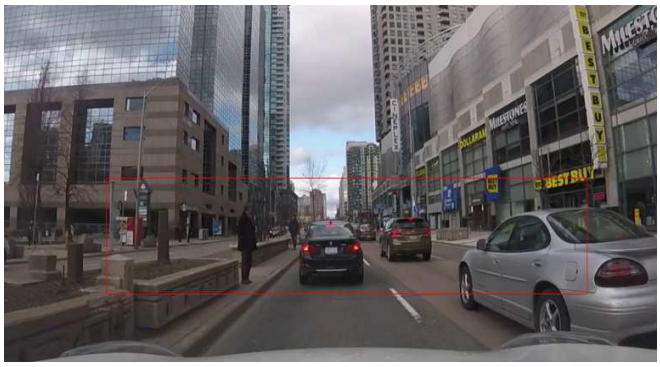


Fig. 2 Selection of a far-distance region. The red box denotes the far-distance area to be detected.

### B. Super-resolution with GAN

The SR module ultimately converts a small image of a pedestrian into a high-resolution image. The conventional method uses bilinear and bi-cubic interpolation to enlarge the image, but the overall effect is unsatisfactory. In this paper, we used DNNs to address this. The DNNs are convolutional neural networks (CNNs). The DNNs perform three processes. First, the feature map is extracted from a low-resolution image through patch extraction and representation. Second, layers are used to add non-linearity to the feature map extracted in the first step. The final step is reconstruction, in which high-resolution images are extracted from the non-linear feature map.

A GAN is a neural network that generates virtual data. To generate good target data, a generator module is configured and iteratively trained to deceive a discriminator module. Resolution enhancement using the DNNs began with a CNN model that used a convolutional layer, followed by a very-deep super-resolution (VDSR) model that used a skip- or residual-connection technique [13]. Subsequently, SRGAN models using generated adversarial neural networks have been developed. The SRGAN model is a GAN-based model that applies the skip-connection method (which is used for CNN and VDSR) to the generator and effectively integrates older technology with new technology [13]. This paper uses SRGAN to magnify small images of pedestrians observed from a distance.

### C. Pedestrian Pose Estimation with Confidence Maps and Affinity Fields

It is difficult to match due to the connection with the existing bottom-up detected body parts, and there are problems such as reduced accuracy and increased calculation amount due to the increase in combinations, and it is intermediate between the body parts. Methods for adding positional information, such as adding additional points, have been proposed. However, their utility is limited without directional information. Cao encodes both the position of body parts and the relationships among them in 2D vector fields that comprise Part Affinity Fields, creating a filter that can be encoded into a vector [3]. This enables pose estimation using the bottom-up approach, which first identifies joints and then performs pose estimation for several people. Accuracy and speed are improved compared to the top-down method, which detects several people and then performs pose estimation for each individual. The figure below shows the encoding flow between each body part for each channel.

In Figure 3, F inputs the image to vgg-19 [25] and converts it using the intermediate layer of feature values. A confidence map is predicted by Branch 1, and Part Affinity Fields are predicted by Branch 2. Branches 1 and 2 are executed repeatedly (stages 1–6), with the input for each stage consisting of the output from the previous stage. The input F from Stage 1 is concatenated and used in all stages. The error (loss function) is calculated for each stage, improving accuracy with each repetition. In this study, confidence maps and affinity fields are calculated for an image of a pedestrian in the far distance, which was expanded by the previous module, and pose estimation is then performed.

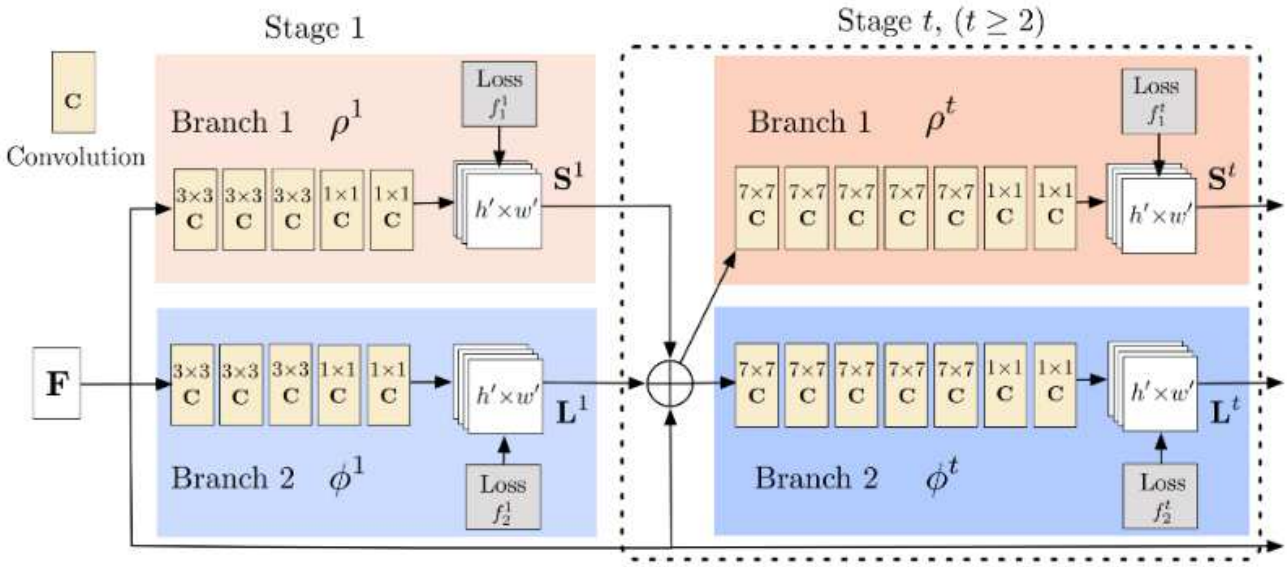



Fig. 3 The architecture of the two-branch multi-stage CNN for human pose estimation [3]. The first branch estimates confidence maps $\boldsymbol{S^t}$, and the second branch estimates Part of Affinity Fields $\boldsymbol{L^t}$.

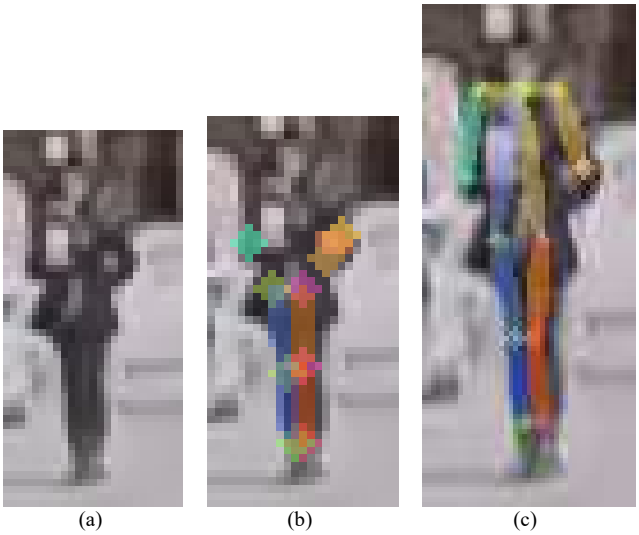
## III. Results and Discussion

The dataset used in this paper is the Joint Attention in Autonomous Driving (JAAD) pedestrian data set [27] - [30]. The JAAD includes a data set for autonomous driving and a data set for pedestrian and driver behavior at crossing points. Figure 4 is an example of a JAAD pedestrian dataset. It provides 346 short, annotated video clips (each 5–10 seconds long) extracted from over 240 hours of driving footage. The video was filmed in many parts of North America and Eastern Europe and includes typical scenes of urban driving in various weather conditions. A mono camera filmed the video, and the size of the test images is 1,920 × 1,080 pixels.

We used the data from the 346 short driving clips and tested our method on a total of 792 pedestrians in the far distance. The annotation data provided by JAAD were used to carry out comparative experiments. Annotation data do not include distance data. Therefore, we defined the targets as pedestrians of a minimum size (as indicated by data labels). Annotation

data also included human pose data, necessitating another evaluation method. In this paper, we assumed that the pose estimation data were accurate and that poses were evenly distributed over a certain region of the image box.

Fig. 4 Example of a JAAD pedestrian dataset. The image shows a pedestrian observed while driving in the city. Annotation data are obtained from public data of the city.



(a) (b) (c)

Fig. 5 Results of pose estimation of a female pedestrian in the far distance. (A) The image box for a pedestrian in the far distance. (B) This is an example of inaccurate pose estimation resulting from the use of an unenlarged image. (C) This is an example of accurate pose estimation after enlarging the image using SRGAN.

Figure 5 shows the results of applying the method proposed in this paper. Figure 5(a) shows a pedestrian detected in the far distance. The size of the image box was 25 × 55 pixels. In Figure 5(b), the upper body parts were inaccurately estimated as a result of using Figure 5(a) for pose estimation without enlargement. Figure 5(c) shows the results of the method proposed in this paper. The image was expanded to 50 × 110 pixels using SRGAN. The coordinates of the image will be re-reduced for use.

(a) (b) (c)

Fig. 6 Results of pose estimation of a male pedestrian in the far distance. (A) The image box for a pedestrian in the far distance. (B) This is an example of inaccurate pose estimation resulting from the use of an unenlarged image. (C) This is an example of accurate pose estimation after enlarging the image using SRGAN.

Figure 6 shows the results of applying the method proposed in this paper to a male pedestrian. Figure 6(a) shows a pedestrian detected in the far distance. The size of the image box was 26 × 73 pixels. In Figure 6(b), the upper body parts are not properly connected because pose estimation was applied to an unenlarged image. Figure 6(c) shows the results of the method proposed in this paper. The image was expanded to 52 × 146 pixels using SRGAN.

In our experiments, 23.4% of pose estimations were successful when using the conventional method for detecting pedestrians at a far distance. With the method proposed in this paper, pose estimation was successful in 69.2% of cases, for a performance improvement of 45.8%. Pose estimation was unsuccessful even when using the proposed method in cases where annotation data indicated that a pedestrian was present, but the body was obscured by a vehicle or another object in the image. When pedestrians were partially hidden, pose estimation could not be performed.

## IV. Conclusion

In this paper, we have proposed a method to solve the pose estimation problem for pedestrians in the far distance. First, DNNs were applied to estimate the vanishing point. Next, this information was used to select the far-distance region, after which DNNs were used to detect pedestrians. The target image was then enlarged, and SRGAN was used to increase image resolution. Finally, we have applied DNNs to estimate the pose of the pedestrian. Experimental results demonstrate the success of this method's sequential application of DNNs. In future work, we intend to simplify the steps conducted by the DNNs, and experiment with merging and other deformations of the network structure.

## ACKNOWLEDGMENT

Hanyang Women's University supported this research for research funding [grant number: 2021-2-011].

## REFERENCES

[1] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4903-4911, 2017.

[2] H. Wang, R. A. Güler, I. Kokkinos, G. Papandreou, and S. Zafeiriou, "BLSM: A bone-level skinned model of the human mesh," *In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 1-17, 2020.

[3] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291-7299, 2017.

[4] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," *IEEE transactions on pattern analysis and machine intelligence, 43(1)*, pp. 172-186, 2021.

[5] Z. Cao, H. Gao, K. Mangalam, Q. Z. Cai, M. Vo, and J. Malik, "Long-term human motion prediction with scene context," *In Computer Vision–ECCV 2020: 16th European Conference*, pp. 387-404, 2020.

[6] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," *IEEE transactions on pattern analysis and machine intelligence, 43(1)*, pp. 172-186, 2021.

[7] H. S. Fang, S. Xie, Y. W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 2334-2343, 2017.

[8] J. Li, C. Wang, H. Zhu, Y. Mao, H. S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10863-10872, 2019.

[9] H. S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[10] J. Sun, Y. Li, L. Chai, H. S. Fang, Y. L. Li, and C. Lu, "Human trajectory prediction with momentary observation," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6467-6476, 2022.

[11] H. S. Fang, Y. Xie, D. Shao, Y. L. Li, and C. Lu, "DecAug: augmenting HOI detection via decomposition," *In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 2*, pp. 1300-1308, 2021.

[12] H. S. Fang, Y. Xie, D. Shao, and C. Lu, "Dirv: Dense interaction region voting for end-to-end human-object interaction detection," *In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 2*, pp. 1291-1299, 2021.

[13] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681-4690, 2017.

[14] J. M. Wolterink, K. Kamnitsas, C. Ledig, and I. Išgum, "Deep learning: Generative adversarial networks and adversarial methods," *In Handbook of Medical Image Computing and Computer Assisted Intervention*, pp. 547-574, 2020.

[15] C. Rockwell, D. F. Fouhey, and J. Johnson, "Pixelsynth: Generating a 3d-consistent experience from a single image," *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14104-14113, 2021.

[16] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11977-11986, 2019.

[17] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874-1883, 2016.

[18] C. Ouyang, J. Schlemper, C. Biffi, G. Seegoolam, J. Caballero, A. N. Price, and D. Rueckert, "Generalising deep learning MRI reconstruction across different domains," *arXiv preprint arXiv:1902.10815*, 2019.

[19] S. Park, J. Yoo, D. Cho, J. Kim, and T. H. Kim, "Fast adaptation to super-resolution networks via meta-learning," *In Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XXVII 16*, pp. 754-769, 2020.

[20] S. Lee, D. Cho, J. Kim, and T. H. Kim, "Restore from restored: Video restoration with pseudo clean video," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3537-3546, 2021.

[21] S. Lee, J. Kim, J. S. Yoon, S. Shin, O. Bailo, N. Kim, and I. S. Kweon, "Vpgnet: Vanishing point guided network for lane and road marking detection and recognition," *In Proceedings of the IEEE international conference on computer vision*, pp. 1947-1955, 2017.

[22] G. Chen, K. Chen, L. Zhang, L. Zhang, and A. Knoll, "VCANet: Vanishing-point-guided context-aware network for small road object detection," *Automotive Innovation, 4*, pp. 400-412, 2021.

[23] X. Li, L. Zhu, Z. Yu, B. Guo, and Y. Wan, "Vanishing point detection and rail segmentation based on deep multi-task learning," *IEEE Access, 8*, pp. 163015-163025, 2020.

[24] W. Wang, P. Lu, X. Peng, W. Yin, and Z. Zhao, "RLSCNet: A Residual Line-Shaped Convolutional Network for Vanishing Point Detection," *In MultiMedia Modeling: 29th International Conference, MMM 2023*, pp. 103-114, 2023.

[25] G. Welch and G. Bishop, *An introduction to the Kalman filter*, 1995.

[26] M. Khodarahmi and V. Maihami, "A review on Kalman filter models," *Archives of Computational Methods in Engineering, 30(1)*, pp. 727-747, 2023.

[27] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," *In Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 206-213, 2017.

[28] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1258-1268, 2021.

[29] B. Liu, E. Adeli, Z. Cao, K. H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters, 5(2)*, pp. 3485-3492, 2020.

[30] B. Yang, W. Zhan, P. Wang, C. Chan, Y. Cai, and N. Wang, "Crossing or not? Context-based recognition of pedestrian crossing intention in the urban environment," *IEEE Transactions on Intelligent Transportation Systems, 23(6)*, pp. 5338-5349, 2021.