SynPoses: Generating Virtual Dataset for Pedestrian Detection in Corner Cases

Yunhao Nie, Bo Lu, Qiyuan Chen, Qinghai Miao, *Senior Member, IEEE*, and Yisheng Lv, *Senior Member, IEEE*

***Abstract*—Pedestrian detection based on deep learning methods makes a big hit during these days. The key to achieve excellent results for deep learning-based pedestrian detection methods is a high-quality dataset. There are lots of outstanding datasets for pedestrian detection like EuroCity Persons, CityPersons, Caltech, etc. However, their efforts are not dedicated to the traffic scenarios with pedestrians in various poses. When the state-of-the-art detectors trained on datasets described above encounter the scene with pedestrians in variously complicated poses, the results are mostly worse than expected. In order to alleviate this problem, we propose a framework called SynPoses to create synthetic human with complicated poses in high quality. Experimental results show that, when facing scenarios with human in diverse poses, the performance of detectors trained on augmented dataset outperforms those trained on original dataset.**

***Index Terms*—Pedestrian, detection, augmentation, dataset.**

I. Introduction

IN AUTONOMOUS driving, an excellent perception system is the basis for subsequent decisions of the control system and execution system. Among perception tasks, pedestrian detection is essential to safety in urban traffic scenarios. Whether the autonomous driving perception system can identify all kinds of pedestrians in various scenes is not only about the subsequent performance of the subsequent control system and execution system of intelligent vehicles [1], but also ensure the safety of pedestrians and vehicles in complex traffic scenes [2].

At present, autonomous driving perception systems increasingly rely on various detection and recognition models such as those based on deep learning (e.g., Mask R-CNN [3]), however, deep learning methods need to be trained with large datasets to achieve their expected results. Moreover, when it comes to the corner case detection for autonomous driving. The real-world traffic scene dataset cannot provide complete and sufficient coverage for corner cases in pedestrian recognition in autonomous driving scenarios, and when these corner cases occur, the perception system will not be able to respond effectively, which may further lead to traffic accidents [4]. For example, what is unexpected but reasonable is that some normal poses (e.g., sit, bend over, lay down) are actually relatively rare in the real traffic scene. According to ACP framework of Parallel Intelligence [5], [6], execution in real world can benefit from artificial systems. For this reason, virtual synthetic datasets for covering these scenes and further diffusion from virtual to real traffic scenes are of great value. Therefore, an augmented dataset for pedestrians' poses is needed.

Our work focuses on the method to augment existing dataset within pedestrians in real traffic scene and tries to improve the performance of pedestrian detectors when facing the scenarios of pedestrians in various poses. The main contributions of this paper are as follows.

1. To the best of our knowledge, we are the first to explore a data augmentation method based on real traffic scenarios of pedestrian detection regarding pedestrian poses in the corner case.

2. A framework is proposed to generate effective augmented dataset based on real traffic scenarios, composited with synthetic but realistic pedestrians in diverse and yet reasonable poses.

It is worth pointing out that our augmented dataset based on half size of EuroCity Persons [7] dataset (ECP) can achieve better generalization performance compared to the original one but even on the entire ECP dataset.

II. Related Work

A. Pedestrian Detection Datasets

One of the key tasks in autonomous driving perception systems is pedestrian detection. Training deep learning-based algorithmic models, such as some models for object detection such as pedestrians, requires a large amount of data. An increasing number of datasets based on realistic traffic scenarios have emerged in recent years, such as Caltech [8], KITTI [9], CityPersons [10], EuroCity Persons [7], etc. As shown in Table I, they provide a large number of informative images and annotations of traffic scenes.

However, due to various factors in real scenes, the distribution of diverse poses in traffic scenes has limitations. This

Manuscript received 25 July 2022; revised 13 September 2022; accepted 19 September 2022. Date of publication 3 October 2022; date of current version 12 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant U1811463, and in part by the Open Project of the State Key Laboratory of Management and Control for Complex Systems under Grant 20220117. *(Corresponding author: Qinghai Miao.)*

Yunhao Nie, Bo Lu, Qiyuan Chen, and Yisheng Lv are with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: nieyunhao2020@ia.ac.cn; lubo2021@ia.ac.cn; chenqiyuan2020@ia.ac.cn; yisheng.lv@ia.ac.cn).

Qinghai Miao is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: miaoqh@ucas.ac.cn).

Digital Object Identifier 10.1109/JRFID.2022.3211285

TABLE I
COMPARISON OF PEDESTRIAN DATASETS

	Caltech	KITTI	CityPersons	EuroCity Persons
images(day/night)	249884/-	14999/-	5000/-	40219/7117
Pedestrian	289395/	9400/-	31514/-	183182/35323

further leads to the fact [7] that the aspect ratio of most pedestrians after labeling lies in the range of [0.28, 0.42]. Even the annotation strategy of CityPersons use a fixed aspect ratio to label all the pedestrians. The consequence of this is that on the one hand, it is not reasonable for a few existing pedestrians with diverse poses, and on the other hand, it still cannot cover pedestrians in various pose scenarios.

B. Frameworks for Data Augmentation

The data augmentation method can not only reduce the overfitting problem of deep learning algorithms, but also expand the coverage of specific case scenarios [11]. Lv et al. [12] proposed to use generative adversarial network (GAN) method for data augmentation in the traffic scene. Liu et al. [13] further proposed "APGAN" to increase training data coverage, to perform data augmentation on the existing pedestrian detection model.

At the same time, great progress has also been made in the fields of human action pose synthesis and 3D human reconstruction. Loper et al. [14] proposed a learning model "SMPL" for human shape and pose-related shape changes, which can provide more refined rendering results than other models at the time, in addition to ensuring compatibility with graphics rendering workflows. Later, Pavlakos et al. present Vposer [15], a learning based variational human pose prior trained from a large dataset of human poses represented as SMPL bodies.

Furthermore, Vobecký et al. [16] proposed a method to enhance the portrait dataset by synthesizing people with adjustable pose and appearance into a real urban background.It demonstrates that neural networks of various model complexities designed for multiple tasks benefit from artificially generated samples, especially when the data distribution can be controlled. Zhang et al. [17] explore the method compositional data augementation via learning object placement by inpainting. These works make a big step to the data augmentation, however, they still don't cover the scenes of pedestrians with diverse poses.

III. THE METHOD OF DATA GENERATION

We propose a framework (as show in Fig. 1) for generating synthetic data to achieve the augmentation on original real traffic data. In the synthesis phase of the data, we mainly focus on and address three problems as following:

1. How to synthesize realistic pedestrians and ensure their poses' plausibility?

2. How to composite the synthetic pedestrian into the real traffic scene at reasonably position?

3. Whether the augmented dataset effectively improve pedestrian detector?

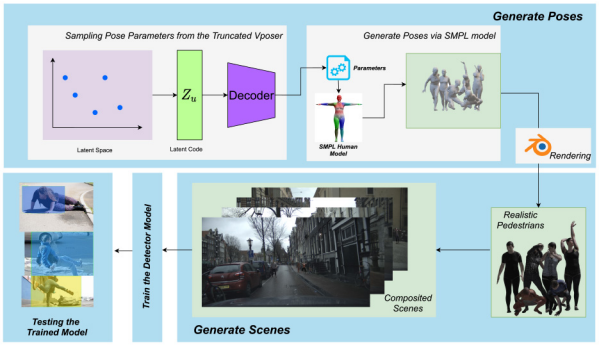

Fig. 1. The overall framework of data generation.

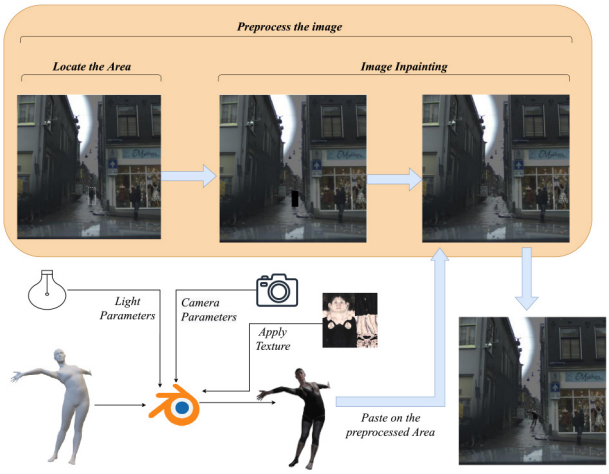

Fig. 2. Details on the pipeline of compositing synthetic pedestrian.

A. The Generation of Pedestrians' Poses

For the aim of generating the synthetic pedestrians and ensure their poses' plausibility, we choose to truncate the VAE model of the VPoser because of its prior over body pose, by sampling from the latent space $Z \in \mathbb{R}^{32}$. Furthermore, we set a reasonable parameter for Gaussian distribution to sample a large number of human poses, where μ is the mean and σ the standard deviation.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

After that, when it comes to ensure the fidelity of synthetic pedestrians, inspired by Varol et al. [18], we choose a random strategy to apply the textures to a synthetic human. At the same time, we also set fixed parameters to render the light and the position of camera in the Blender,[1] as shown in Fig. 2.

B. Embedding Synthetic Pedestrian in Real Scenes

Firstly, since the augmentation is based on the real dataset collected from the car-oriented perspective, we choose the Eurocity Persons dataset because of its rich labels including bounding box and pedestrians' orientation in abundant traffic scenarios.

[1]https://www.blender.org

TABLE II
THE TEST RESULTS ON OCHUMAN DATASET. NOTE: *AP* STANDS FOR AP AT IOU=.50:.05:.95 (PRIMARY CHALLENGE METRIC), $AP^{IoU=.50}$ STANDS FOR AP AT IOU=.50 (PASCAL VOC METRIC), $AP^{IoU=.75}$ STANDS FOR AP AT IOU=.75 (STRICT METRIC), $AR^{max=100}$ STANDS FOR AR GIVEN 100 DETECTIONS PER IMAGE, $AR^{max=10}$ STANDS FOR AR GIVEN 10 DETECTIONS PER IMAGE, $AR^{max=1}$ STANDS FOR AR GIVEN 1 DETECTIONS PER IMAGE

	Test on OCHuman Dataset					
	AP	$AP^{IoU=.50}$	$AP^{IoU=.75}$	$AR^{max=100}$	$AR^{max=10}$	$AR^{max=1}$
Half ECP Dataset	4.1%	12.4%	2.2%	41.2%	24.6%	7.0%
ECP Dataset	13.5%	33.1%	9.2%	57.5%	44.9%	16.2%
Augmented Half ECP Dataset	**18.9%**	**40.8%**	**15.2%**	**60.4%**	**47.9%**	**21.5%**

(a) (b) (c)

Fig. 3. Comparison with different embedding strategies: (a) the original scene, (b) the composited scene which only considered the positional semantics: synthetic pedestrians are labeled with red bounding box, (c) the composited scene under our embedding strategy: synthetic pedestrians are labeled with blue bounding box.

(a) Examples of detection results by model trained on Augmented Half ECP Dataset

(b) Examples of detection results by model trained on Half ECP Dataset

Fig. 4. The detection result of examples.

Secondly, we draw the inspiration from Ghiasi et al. [19] and Zhang et al. [17]. As shown in Fig. 3(b), on the one hand, a simple copy-paste is done by referring the existing pedestrians; on the other hand, the original pedestrian is removed if it does not fit the paste, and the image is restored using the "co-modulated" generative adversarial network method proposed by Zhao et al. [20]. After the image-inpainting process, then the synthetic pedestrian is embedded into the processed area. What is more, as shown in Fig. 3(b), it is note-worthy that the occlusion, density, positional reasonableness and semantic reasonableness are considered in the strategy of embedding by taking into account the positional and occlusion relationships between synthetic pedestrians and real pedestrians.

Finally, through the methods and steps described above, we generate the augmented dataset based on a half size of ECP dataset. In addition, we also compared with other data augmentation methods in various aspects, as shown in the Table III.

TABLE III
THE COMPARISON OF AUGMENTED METHODS

Method	***Inpainting***	***Pose Diversity***	***Fine Texture***
APGAN [13]	✗	✗	✓
PlaceNet [17]	✓	✗	✗
DummyNet [16]	✗	✗	✗
SynPoses (Ours)	✓	✓	✓

C. Experiment

For the aim of testing the generalization performance of each datasets, we choose the state-of-the-art detector Pedestron [21] on ECP dataset for testing. Furthermore, we choose the Cascade Mask-R-CNN as the baseline. What is more, we choose the Eurocity Persons dataset as the benchmark dataset for training. To show the superior performance of our augmentation method, we choose half of the ECP dataset for the data augmentation as the training dataset to compare with the original one. Under the condition that fixed with the same training hyperparameters, we train the Cascade Mask-R-CNN model for 20 epochs on the ECP, half of ECP, and augmented half of ECP dataset separately.

As for the test dataset, we choose the OCHuman [22] due to its diverse poses and abundant scenes. It contains 13360 elaborately annotated human instances within 5081 images with comprehensive labels including bounding-box, humans pose. OCHuman is the most complex and challenging dataset regarding human due to average 0.573 MaxIoU of each person.

Unless otherwise specified, the experimental results in the Table II are based on the conditions of detection for all areas and given 100 detections per image. Regarding the evaluation metrics, we choose the COCO evaluation metrics [23]. AP and AR are averaged over multiple Intersection over Union (IoU) values. Specifically 10 IoU thresholds of.50:.05:.95 is used in our experiment setting due to averaging over IoUs rewards detectors with better localization.

We can draw conclusions from the Table II and the Fig. 4 that our augmented dataset deal way better than the half of

original dataset and even the entire original dataset when faced with the scenarios within complex pedestrian poses.

IV. Conclusion

To sum up, we propose a data augmentation pipeline based on real traffic scenarios by implementing an augmentation method trying to cover corner cases of training pedestrian poses for autonomous driving. The aim is training pedestrian detectors through the augmented dataset, so as to improve the performance of pedestrian detection when facing hard samples of various poses. We choose ECP as the benchmark dataset to augment and train baseline model on the original ECP and the augmented one. Further, we test directly on the challenging OCHuman dataset. Experimental results show that the detectors trained on the augmented dataset outperform those trained on the original dataset in these scenarios which included occluded and various-poses pedestrians. It turns out that our augmented dataset is better than the original one and effectively improves the detection of pedestrians in various poses.

References

[1] J. E. Hoffmann, H. G. Tosso, M. M. D. Santos, J. F. Justo, A. W. Malik, and A. U. Rahman, "Real-time adaptive object detection and tracking for autonomous vehicles," *IEEE Trans. Intell. Veh.*, vol. 6, no. 3, pp. 450–459, Sep. 2020.

[2] M. Schutera, M. Hussein, J. Abhau, R. Mikut, and M. Reischl, "Night-to-day: Online image-to-image translation for object detection within autonomous driving by night," *IEEE Trans. Intell. Veh.*, vol. 6, no. 3, pp. 480–489, Sep. 2021.

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

[4] V. Fors, B. Olofsson, and L. Nielsen, "Autonomous wary collision avoidance," *IEEE Trans. Intell. Veh.*, vol. 6, no. 2, pp. 353–365, Jun. 2021.

[5] F.-Y. Wang, "Parallel system methods for management and control of complex systems," *Control Decis.*, vol. 19, pp. 485–489, Jan. 2004.

[6] F.-Y. Wang, "Artificial societies, computational experiments, and parallel systems a discussion on computational theory of complex social-economic systems," *Complex Syst. Complexity Sci.*, vol. 1, no. 4, pp. 25–35, 2004.

[7] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "The EuroCity persons Dataset: A novel benchmark for object detection," 2018, *arXiv:1805.07193*.

[8] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 304–311.

[9] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[10] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3213–3221.

[11] X. Li, K. Wang, Y. Tian, L. Yan, F. Deng, and F.-Y. Wang, "The ParallelEye dataset: A large collection of virtual images for traffic vision research," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2072–2084, Jun. 2019.

[12] Y. Lv, Y. Chen, L. Li, and F.-Y. Wang, "Generative adversarial networks for parallel transportation systems," *IEEE Intell. Transp. Syst. Mag.*, vol. 10, no. 3, pp. 4–10, Jun. 2018.

[13] S. Liu et al., "A novel data augmentation scheme for pedestrian detection with attribute preserving GAN," *Neurocomputing*, vol. 401, pp. 123–132, Aug. 2020.

[14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, 2015.

[15] G. Pavlakos et al., "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10975–10985.

[16] A. Vobeckỳ, D. Hurych, M. Uřičář, P. Pérez, and J. Sivic, "Artificial dummies for urban dataset augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2692–2700.

[17] L. Zhang, T. Wen, J. Min, J. Wang, D. Han, and J. Shi, "Learning object placement by inpainting for compositional data augmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 566–581.

[18] G. Varol et al., "Learning from synthetic humans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 109–117.

[19] G. Ghiasi et al., "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2918–2928.

[20] S. Zhao et al., "Large scale image completion via co-modulated generative adversarial networks," 2021, *arXiv:2103.10428*.

[21] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable pedestrian detection: The elephant in the room," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 11328–11337.

[22] S.-H. Zhang et al., "Pose2Seg: Detection free human instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 889–898.

[23] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.