Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 21(1): 98-102, Mar. 2023 Regular paper

Human Detection using Real-virtual Augmented Dataset

Jongmin Lee[1], Yongwan Kim[2], Jinsung Choi[2], Ki-Hong Kim[2], and Daehwan Kim[1]*

[1]School of IT Convergence, University of Ulsan, 44610, South Korea
[2]VR/AR Content Research Section, Communications & Media Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), 34129, South Korea

Abstract

This paper presents a study on how augmenting semi-synthetic image data improves the performance of human detection algorithms. In the field of object detection, securing a high-quality data set plays the most important role in training deep learning algorithms. Recently, the acquisition of real image data has become time consuming and expensive; therefore, research using synthesized data has been conducted. Synthetic data haves the advantage of being able to generate a vast amount of data and accurately label it. However, the utility of synthetic data in human detection has not yet been demonstrated. Therefore, we use You Only Look Once (YOLO), the object detection algorithm most commonly used, to experimentally analyze the effect of synthetic data augmentation on human detection performance. As a result of training YOLO using the Penn-Fudan dataset, it was shown that the YOLO network model trained on a dataset augmented with synthetic data provided high-performance results in terms of the Precision-Recall Curve and F1-Confidence Curve.

Index Terms: Data augmentation, Human detection, Semi-synthetic data, YOLO etc.

I. INTRODUCTION

One of the most important factors for improving the accuracy of a deep learning model is securing a high-quality dataset. Because it is directly proportional to the amount of data used for training. Building a dataset requires considerable time and money, so many techniques for data augmentation have been used [1].

Basic image processing techniques such as cropping, rotation, scale, shearing, flipping or reflection, and translation are widely used for image data augmentation [2]. Recently, techniques that generate images by combining objects and backgrounds from different images or by using autoencoder (AE) or generative adversarial network (GAN) techniques have been used. These data augmentation methods [3] contribute significantly to improving the performance of deep learning models. However, because these methods are synthesized or created using existing data sets, there are limitations in increasing the amount of data in a completely new form. In addition, there are limitations in obtaining real image data or labeling vast amounts of data.

To overcome these limitations in securing real image data, various studies [4-6] have been conducted to generate synthetic data and using them for training deep-learning models. There are two major types of synthetic data: fully synthetic and semi-synthetic.

Fully synthetic data were generated by graphical modeling of all backgrounds and objects. It can generate large amounts of data and simplify labeling. Owing to the performance of graphic modeling at the level of photorealism, it has begun to replace real data. However, it is still used only as an auxiliary data form because of the difference between the fundamental data form and the real data.

Semi-synthetic data were created by properly mixing real

Received 4 December 2022, **Revised** 16 February 2023, **Accepted** 17 February 2023
***Corresponding Author** Daehwan Kim (E-mail: daehwankim@ulsan.ac.kr, Tel: +82-52-259-2215)
School of IT Convergence, University of Ulsan, 44610, Republic of Korea

Open Access https://doi.org/10.56977/jicce.2023.21.1.98 print ISSN: 2234-8255 online ISSN: 2234-8883

and virtual data using graphic/video editing tools. Generally, real-background/virtual objects or virtual-background/real objects are mixed and used. Many studies have used semi-synthetic datasets to improve object detection and recognition performance.

In this paper, we introduce a study on how the use of semi-synthetic data can improve human detection performance.

II. SEMI-SYNTHETIC HUMAN DATA GENERATION

We used the Unity game engine [7], to create semi-synthetic human data. Semi-synthetic human data were created by synthesizing a real image background and virtual human model. The background image was used by deleting the human area from the Penn-Fudan dataset [8], and the virtual human models were used by downloading free models from the Unity Store. To generate human-detection image data under the same conditions, the human region was removed and a virtual human model was synthesized in a similar location. Figure 1 shows an example of the removal of human regions from an image in the Penn-Fudan dataset.

Fig. 1. An example of an original image and its human-regions removed image.

The Clean-Up Pictures tool [9] was used for human region removal. If an image with a person is input, an image with the person removed can be obtained. Subsequently, using the Unity engine, a virtual human model was synthesized on the image from which people were removed. Figure 2 shows an example of a synthesized image of a virtual human model.

III. OBJECT DETECTION ALGORITHM: YOLOv5

We used You Only Look Once (YOLO) [10] as an object detection algorithm to test the augmentation effect of semi-synthetic human data. The YOLO series provides high speed and accuracy for real-time object detection. Many studies have used the YOLO algorithm, which is popular owing to its ease of use. In addition, it is one of the best object detection algorithms available.

Fig. 2. An example of the synthesized image of virtual human models

The old YOLO algorithm is a one-stage algorithm that detects object regions using a regressive network structure. Although this provides high speed, the detection performance for small objects is moderately poor. The latest version of YOLOv5 improves the detection performance for small objects. The YOLOv5 algorithm uses the Cross Stage Partial Network (CSPNet) [11] and the Full Convolutional One-Stage (FCOS) method [12]. They used a cross-stage feature fusion strategy to propagate gradients efficiently. This improves the efficiency and accuracy of object detection.

IV. COMPARISON OF HUMAN DETECTION PERFORMANCE BASED ON SEMI-SYNTHETIC DATA AUGMENTATION

To compare human detection performance based on semi-synthetic data augmentation, the YOLO algorithm was trained in two different ways and the Penn-Fudan was used for human detection training. The first was trained using the full real dataset, and the second was trained by replacing 20% of the images with semi-synthetic data in the full real dataset.

A total of 170 images were used as training images. The first detection network was trained using all 170 real images, and the second was trained using semi-synthetic data from 34 images (20% of the 170 images).

Fourteen virtual human models were used for the semi-synthetic data. We selected random images from the Penn-Fudan dataset and synthesized between one and three virtual human models per image to create semi-synthetic data.

All other conditions were the same. The batch size was 16, and the number of epochs was 20 for training.

To accurately compare human detection performance, the

pedestrian dataset [13] was used as the test dataset. The reason for using separate training and test datasets was to make a complete universal performance comparison. Figure 3 shows an example image of the pedestrian dataset used to test the human detection performance.

Fig. 3. Example images from the Pedestrian dataset.

Two quantitative criteria were used to verify the performance of the YOLO network model trained on each dataset. These are the precision-recall and the F1-confidence curve.

Figures 4 and 5 show the precision-recall curve graphs for testing the YOLO model trained using full real human data and semi-synthetic human data, respectively. Both models showed values above 0.9, but the model based on semi-synthetic data showed a slightly better detection performance.

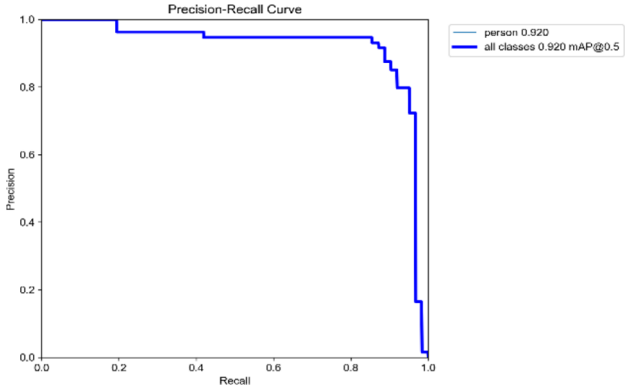

Fig. 4. The Precision-Recall Curve graph of YOLO model trained with full real human data.

Figures 6 and 7 show graphs of the F1-Confidence Curve for each model test. In this graph, the model using semi-synthetic data showed relatively better performance.

This experimentally confirms that constructing a network using semi-synthetic data helps improve the general-purpose performance of human detection.

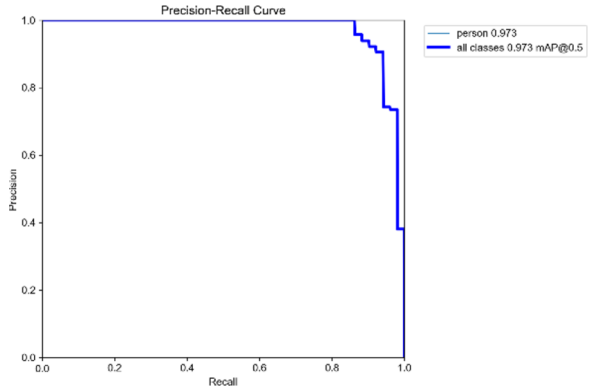

Fig. 5. The Precision-Recall Curve graph of YOLO model trained with semi-synthetic human data.

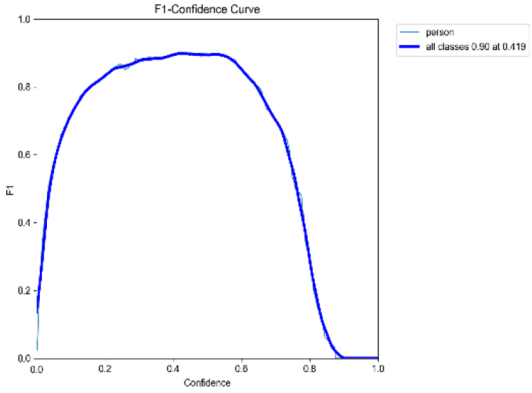

Fig. 6. The F1-Confidence Curve graph of YOLO model trained with full real human data.

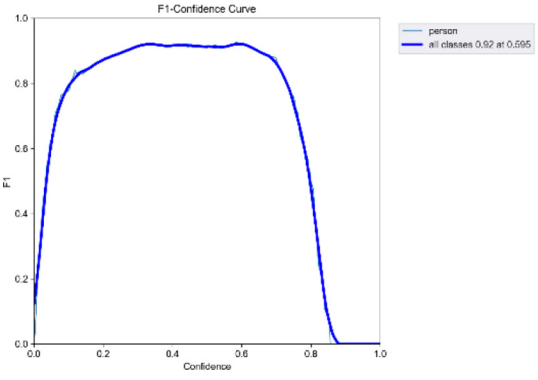

Fig. 7. The F1-Confidence Curve graph of YOLO model trained with semi-synthetic human data.

Finally, the change in the confidence value was investigated using human detection image results. Figure 8 and Figure 9 show the human detection result images of the pedestrian dataset of the network trained with full real human data and semi-synthetic human data, respectively. Both the models exhibited excellent human detection performance.

However, the confidence values for detecting the same human were slightly different. The confidence value of the

network model trained using semi-synthetic data was slightly low. It is estimated that this is the result of learning via a 3D graphic model rather than an actual image.

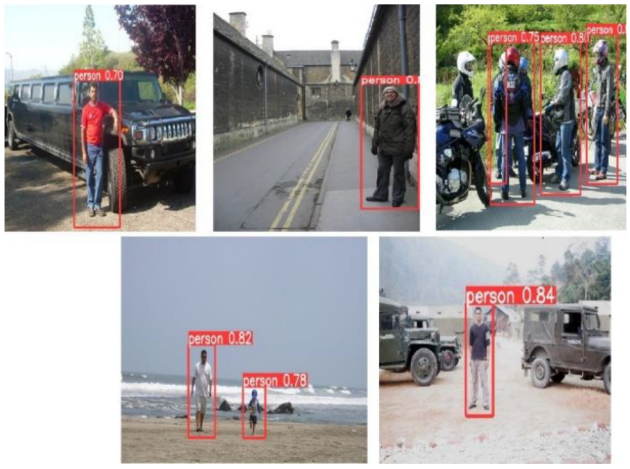

Fig. 8. Image examples of human detection test results of the YOLO model trained with full real human data.

Fig. 9. Image examples of human detection test results of the YOLO model trained with semi-synthetic human data.

V. CONCLUSION

Recently, the use of synthetic data to train deep-learning network models is being investigated. In this study, we investigated the effect of semi-synthetic data augmentation on the human detection performance using the YOLOv5 algorithm. Through comparison with the network trained with a complete real data set, it was experimentally confirmed that the general-purpose performance was relatively high in terms of the Precision-Recall Curve and F1-Confidence Curve. We believe that training deep- learning models using synthetic data is one way to save time and money. As the graphic rendering performance improves in the future, the utilization of synthetic data is expected to increase.

ACKNOWLEDGMENTS

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sprots and Tourism in 2022 (Project Name: Development of Virtual Reality Performance Platform Supporting Multiuser Participation and Realtime Interaction, Project Number: R2021040046, Contribution Rate: 100%)

REFERENCES

[1] I. Joshi, M. Grimmer, C. Rathgeb, C. Busch, F. Bremond, and A. Dantcheva, "Synthetic data in human analysis: A survey," *arXiv preprint arXiv:2208.09191*, Aug. 2022. DOI: 10.48550/arXiv.2208.09191.

[2] H, Kim, D, Kim, J, Kim, and S, Im, "Data augmentation scheme for semi-supervised video object segmentation," *Journal of Broadcast Engineering*, vol. 27, no. 1, pp. 13-19, 2022.

[3] K. Man and J. Chachi, "A Review of Synthetic Image Data and Its Use in Computer Vision," *Journal of Imaging*, vol. 8, no.11, p. 810, Nov. 2022. DOI: 10.3390/jimaging8110310.

[4] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Honolulu, USA. pp. 109-117, 2017. DOI: 10.1109/CVPR.2017.492.

[5] J. Mu, W. Qiu, G. D. Hanger, and A. L. Yuille, "Learning from synthetic animals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 12386-12395, 2020. DOI: 10.1109/CVPR42600.2020.01240.

[6] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Pixel-wise crowd understanding via synthetic data," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 225-245, Jan. 2021. DOI: 10.1007/s11263-020-01365-4.

[7] Unity. [Online] Available: https://www.unity.com.

[8] Penn-Fudan Database for Pedestrian Detection and Segmentation. [Online] Available: https://www.cis.upenn.edu/~jshi/ped_html/.

[9] Remove any unwanted object, defect, people of text you're your pictures in seconds, cleanup.pictures. [Online] Available: https://cleanup.pictures/

[10] G. Jocher, Code. [Online] Available: https://github.com/ultralytics/yolov5.

[11] C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, USA, pp. 390-391, 2020. DOI: 10.1109/CVPRW50498.2020.00203.

[12] Z, Tian, C, Shen, H. Chen, and T, He, "FCOS: full convolutional one-stage object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea, pp. 9627-9636, 2019. DOI: 10.1109/ICCV.2019.00972.

[13] Pedestrian Detection Data set, Kaggle. [Online] Available: https://www.kaggle.com/datasets/karthika95/pedestrian-detection.

Jong-Min Lee

He has been majoring in Computer Science at Ulsan University, Korea since 2018. His research interest includes Computer Vision, Data Augmentation, Synthetic Data, and AI.

Yongwan Kim

He completed a B.S. degree (1996) in Electronics Engineering at Inha University, a M.S.E. (1998) in Information and Communications Engineering at GIST, and a Ph.D. (2014) in Computer Science at Korea Advance Institute of Science and Technology (KAIST), Korea. He joined the Electronics and Telecommunications Research Institute (ETRI) in 1998 and has been working as a principal researcher of the Virtual Reality Research Team since then. His research interests include virtual reality, haptics, and human computer interaction.

Jin Sung Choi

He completed an M.S. degree in electrical engineering from Kyungpook National University, Korea, in 1994. He is currently a principal researcher at the Electronics and Telecommunications Research Institute. His research interests include Human-computer Interface, Virtual Reality, Augmented Reality, and empathic computing.

Ki-Hong Kim

He completed his Ph.D. in electrical engineering from Korea Advanced Institute of Science and Technology, Korea in 2007. Since 1996, he has working in the Digital Content Research Division in Electronics and Telecommunications Research Institute (ETRI), where he also works as a principal member of the engineering staff. His main research interests include VR/AR, and Brain-Computer Interface and Speech recognition.

Daehwan Kim

He completed his Ph.D. in computer science and engineering from Pohang University of Science and Technology in 2011. He is currently a professor at the School of IT Convergence at the University of Ulsan. His main research interests include computer vision, AI, deep learning, and VR/AR.