Age Shouldn't Matter: Toward More Accurate Pedestrian Detection via Self-Training

Shunsuke Kogure[1,3], Kai Watabe[2,3], Ryosuke Yamada[2,3], Yoshimitsu Aoki[1], Akio Nakamura[2], Hirokatsu Kataoka[3],

[1]Keio University
[2]Tokyo Denki University
[3]National Institute of Advanced Industrial Science and Technology (AIST)
[1]skogure@aoki-medialab.jp, aoki@elec.keio.ac.jp [2]{watabe.k, yamada.r}@is.fr.dendai.ac.jp, nkmr-a@cck.dendai.ac.jp
[3]hirokatsu.kataoka@aist.go.jp

Abstract

Why is there a the disparity in the miss rates of pedestrian detection between different age attributes? In this study, we propose to (i) improve the accuracy of pedestrian detection using our pre-trained model and (ii) explore the causes of this disparity. In order to improve detection accuracy, we extend a pedestrian detection pre-training dataset, the Weakly Supervised Pedestrian Dataset (WSPD), by means of self-training, to construct our Self-Trained Person Dataset (STPD). Moreover, we hypothesise the cause of the miss rate as being due to three biases: 1) the apparent bias towards "adults" versus "children," 2) the quantity of training data bias against "children," and 3) the scale bias of the bounding box. In addition, we constructed an evaluation dataset by manually annotating "adult" and "child" bounding boxes to the INRIA Person Dataset. As a result, we confirm that the miss rate was reduced by up to 0.4% for adults and up to 3.9% for children. In addition, we discuss the impact of the size and appearance of the bounding boxes on the disparity in miss rates and provide an outlook for future research.

1 Introduction

Recent, research has frequently explored approaches to pedestrian detection , which is expected to be applied in various fields. The remarkable progress that has been made in this area is partly due to the large-scale collection of human images from the Web.

However, there are still concerns about the safety of utilizing pedestrian detection in areas such as automated driving. One of these concerns is the disparity in detection rates based on human age and race ; specifically, a disparity in detection rates between "adults" and "children" has been reported when using classical human detection methods. Brando (Brandao 2019) affirmed that the difference in the quantity of adult versus child data in the person detection dataset is a problem that naturally arises from demographics. There are a small number of "children" in the existing pedestrian dataset, which we assume is responsible for a sample bias and a detection rate disparity between "adults" and "children."

In this paper, we constructed our Self-Trained Person Dataset (STPD) by extending the Weakly Supervisied Pedestrian Dataset (WSPD) (Minoguchi et al. 2020) to improve the accuracy of person detection. We study the effect of each age attribute on detection performance using each pre-trained model generated by the WSPD and STPD. The INRIA Person Dataset (Dalal and Triggs 2005) is used to evaluate the detection performance. We re-annotated both the train and test data of the INRIA Person Dataset to rigorously investigate the effect of age on the accuracy of pedestrian detection. For this re-annotation, we added the age attribute and the bounding box (bbox). In this way, we constructed a dataset for pedestrian detection validation with the age attribute. In addition, we study the reason for the disparity in detection rate by age. Specifically, we examine the age gap in the detection rate using three experiments. (i) We clarify whether there is a difference in appearance between "adults" and "children, (ii) We study the impact of data augmentation of children's learning data alone on the missed rate. (iii) Finally, we compare the miss rate for each age attribute when the scale of the input image is changed. Our contributions are as follows:

- The STPD was constructed by extending the pedestrian dataset, WSPD, using self-training.
- In order to rigorously evaluate the detection performance for "adults" versus "children," we constructed a new evaluation dataset.
- The person detector with STPD pre-training reduced the miss rate of "adults" and "children" compared to the detector with WSPD pre-training. Furthermore, we observed a mitigating effect of self-training on the detection rate gap.
- We studied three aspects to investigate the cause of the gap in detection rates by age: (i) the appearance of "adults" and "children," (ii) the quantity of data for "children," and (iii) the scale of the input images.

2 Related Work

2.1 Detector

In recent years, detection approaches to detection have dramatically with the rise of deep neural networks (DNNs). In the literature, a two-step region identifier and DNN-based classification have been proposed (Girshick et al. 2014). The basic approach, called R-CNN, follows three steps when

Dataset	Image	Bounding Box	Class
Pascal VOC	11,530	27,450	20
MS COCO	123,287	896,782	80
OpenImages v5	1,743,042	14,610,229	600
CityPersons	5,000	35,016	2
EuroCity Persons	47,300	238,200	17
Caltech Pedestrian	250,000	350,000	2
WSPD	2,822,421	8,716,461	2
STPD (Ours)	3,461,024	9,739,996	1
FA-INRIA (Ours)	902	2,993	2

Table 1: Comparison of object detection and person detection datasets.

generating bounding boxes: (i) detect areas in the image that may contain objects (region proposal), (ii) extract CNN features from region candidates, and (iii) Classify objects based on the extracted features. Fast R-CNN (Girshick 2015) also generates region proposals, but it is more efficient than R-CNN because Fast R-CNN pools the CNN features corresponding to each region proposal. Faster R-CNN (Ren et al. 2016) adds a region proposal network (RPN) to generate a region proposal directly in the network. Current research focuses on widely divided one-shot detectors, such as single shot multi-box detector (SSD) (Liu et al. 2016) and you look only once (YOLO) (Redmon et al. 2016). Recent works have also focused on high performance detectors, such as M2Det (Zhao et al. 2019), RetinaNet (Lin et al. 2017), and instance segmentation with Mask R-CNN (He et al. 2017). In this paper, we apply SSD as a method of detecting people in a dataset. Here, we use a WSPD pre-trained model for a self-training.

2.2 Pedestrian Detection

In the past decade, approaches to person detection have improved dramatically. Recent work has proposed configurations to improve recognition and localization, including DNNs , semantic segmentation, combined methods, and small image and cloud analysis. However, in order to train these models, it is necessary to prepare a large dataset and fine-tune its architecture (e.g., SSD or M2Det). Wilson et al. tested whether an object detector can correctly detect pedestrians with different skin colors (Wilson, Hoffman, and Morgenstern 2019). In addition, they found that it is problematic to accurately detect children because their miss rate is higher than that of adults (Brandao 2019). In this study, we were able to detect pedestrians more reliably than in previous studies.

3 Self-Training

3.1 Problem

A number of datasets for pedestrian detection have been proposed so far. However, as shown in Table 1, their scale is small compared to those used for object detection. Minoguchi et al. proposed a weakly supervised learning method that eliminates false positives using existing pre-trained models by referring to bounding boxes and SVM

Annotation Type	Images	%
(i) Adult	2,687	53.7
(ii) Children	169	3.4
(iii) Noise	536	10.7
(iv) Multiple	1,608	32.2

Table 2: The age attribute statistics for people in bounding boxes in 5,000 randomly sampled images from the WSPD dataset. The "Noise" label indicates that there is no person in the bounding box, whereas "Multiple" label means that one bounding box contains multiple people. In this paper, images labeled "Multiple" are not considered.

and by constructing a labeled dataset called the Weakly-Supervised Person Dataset (WSPD) (Minoguchi et al. 2020), which far exceeds the scale of previous pedestrian detection datasets. As far as we have investigated, the WSPD is the largest existing pedestrian dataset. They reveal the detection performance of the pre-trained model on that dataset, they don't mention the disparity in the miss rate for each age attribute. Table 2 shows the attribute distribution of some bounding boxes in the WSPD. This distribution is based on our random selection of 5,000 bounding boxes from the WSPD and their classification by attribute. The "Noise" label indicates that there is no person in the bounding box, while the "Multiple" label indicates that the bounding box contains multiple people. Given this, we can see that the existing pedestrian dataset has a large bias in the distribution of the quantity of the data; in particular, the data for children is too limited. Therefore, it is necessary to check whether this bias in the quantity of data contributes to the disparity in detection performance.

3.2 Solution

As mentioned earlier, we can see that the WSPD contains the largest number of images and bounding boxes among the available person detection datasets. Furthermore, the WSPD contains a wide variety of person images collected from various locations around the world. The semi-automatically collected dataset has millions of bounding boxes, which can be useful for pre-training. We used a WSPD pre-trained model to apply self-training to another dataset to collect high-quality bounding boxes and to investigate the impact of each age attribute on the miss rate. Our self-training pipeline is shown in Figure 1. First, we input images from the Places365 dataset(Zhou et al. 2017) to the SSD, a detector pre-trained with the WSPD, to estimate the location of the bounding box. We assign a pseudo-label of "person" to the predicted bounding box. The determination of the location of the bounding box when generating the pseudo-label is expressed by the following equation:

$$(y', b'_{box}) = D(x; \theta), \tag{1}$$

wherey' and b' represent the predicted values of the object category and bounding box, respectively, and θ represents the learned parameters of the detector. Our self-training approach allows us to automatically extend the

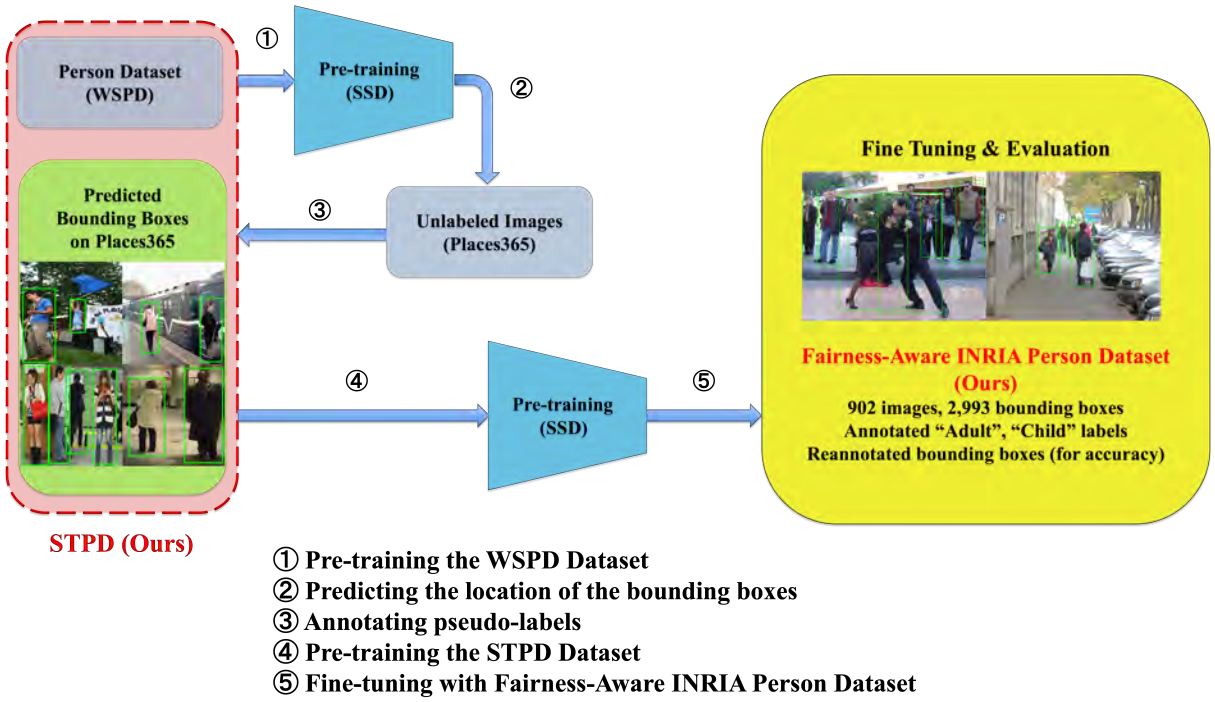

Figure 1: The self-training approach. We used a model pre-trained using the WSPD dataset with the SSD to infer the location of bounding boxes for unlabeled images from the Places365 dataset. We then gave each predicted bounding box a "person" attribute label. By combining these pseudo-labels with the WSPD labels and pre-training them with the SSD, we were able to build a larger model to verify miss rates.

dataset. We refer to the WSPD and the generated pseudo-labeled Places365 dataset together as the Self-Trained Person Dataset (STPD).

Furthermore, we pre-train the constructed STPD and compare its detection performance with the model pre-trained using the WSPD. In order to examine the disparity in the miss rate among age attributes, it is essential to add an age attribute to the bounding box. Then, in order to evaluate the miss rate for each age attribute, we assigned "adult" and "children" labels to the INRIA Person Dataset, which is commonly used for person detection, using the models pre-trained with the WSPD and STPD, respectively. We also re-annotated the location of the bounding box. These two age attributes follow the age categories defined by the Statistics Bureau of the Ministry of Internal Affairs and Communications in Japan for (i) children (0-14 years) and (ii) adults (15 years and older). As a result, we have constructed a pedestrian detection dataset consisting of 902 images and 2,993 bounding boxes for training and evaluation. We named this dataset the Fairness-Aware INRIA Person Dataset (FA-INRIA). An example of the annotations and the breakdown of the dataset attributes are shown in Figure 2 and Table 4, respectively.

Figure 2: Examples of age attribute annotation in Fairness-Aware INRIA Person Dataset (FA-INRIA).

3.3 Experimental Settings

In this paper, we compared the results under the same pre-training conditions. The batch sizes for pre-training the SSD were set to 64, 128, and 256, the number of epochs was set to 10, and the learning rate was set to 0.0005. When we conduct fine-tuning with the FA-INRIA using the pre-trained models on each dataset, the batch size was set to 4, the number of iterations was set to 12,000, and the learning rate was set to 0.0005. Furthermore, the training and test datasets were used with the same configuration as the original INRIA Person Dataset. The experimental settings were described below also conform to these condition.

3.4 Evaluation Metric

We use only the miss rate as an evaluation metric to assess the detection performance for adults and children. In person detection, the relationship between the miss rate and false

Dataset	Batch Size, Epochs	Miss Rate [%] (Adult)	Miss Rate [%] (Children)	Standard Deviation [%]
WSPD	64, 10	13.9	23.1	4.6
	128, 10	13.8	21.2	3.2
	256, 10	13.1	19.2	3.1
STPD (Ours)	64, 10	**13.8**	**19.2**	**2.7**
	128, 10	**13.4**	**19.2**	**2.9**
	256, 10	**13.1**	**17.3**	**2.1**

Table 3: Detection performance comparisons for our **FA-INRIA.** We use the standard deviation to describe the disparity in detection rates between attributes. It is clear that our approach reduces the miss rate for all attributes.

Age	Images	Bounding Boxes
Adult	870	2,672
Children	151	321
All	902	2,993

Table 4: The age attributes in the Fairness-Aware INRIA Person Dataset (FA-INRIA).

positives is often evaluated for each image. However, our goal is to detect all ground truth bounding boxes. Therefore, we calculate the miss rate by examining the breakdown of the age attributes of the bounding boxes that could not be detected. The miss rate M is derived by the following equation:

$$M = 1 - Recall \tag{2}$$

In this paper, we calculated the standard deviation to represent the miss rate disparity among age attributes:

$$v^2 = \frac{1}{n}\sum_{i=1}^{n}(M_i - M)^2, \tag{3}$$

where *n* refers to the number of classes of attributes, which in this study was two ("Adult" and "Children").

3.5 Results

Table 3 shows the miss rate in the FA-INRIA Person Dataset using each of the pre-trained models. Compared to the model pre-trained with the WSPD, the model pre-trained with the STPD reduced the miss rate by up to 0.4% for adults and up to 3.9% for children. In the WSPD pre-trained model, the disparity between the miss rates of adults and children was a maximum standard deviation of 4.6% and a minimum of 3.1%. In contrast, the STPD pre-trained model had a maximum standard deviation of 2.9% and a minimum standard deviation of 2.1%.

Next, the detection results of fine-tuning with the FA-INRIA using the pre-trained detectors on each dataset are shown in Figure 3, illustrating that the STPD pre-trained model is able to detect people that the WSPD pre-trained model misses.

4 Analysis and Discussion

4.1 The relationship between the bias in the quantity of data and the miss rate

In the aforementioned results, we successfully generated a pseudo bounding box containing a person from the Places365 dataset. In Figure 1, we present a visualization of the location of a person's bounding box that was predicted during the process of self-training. This method was implemented based on the success of self-training in object detection (Zoph et al. 2020), and was found to reduce the miss rate for adults and children respectively. Moreover, it is effective in collecting data on pedestrians, regardless of their age attributes, and not only on children, for whom the number of data is small. If the bias in the quantity of data between age attributes is the primary cause of the disparity in detection performance, then it is only the bounding boxes for children that need to be collected more efficiently. However, manual annotation is very costly and impractical. Therefore, we applied data augmentation to the children's bounding boxes in the FA-INRIA training data to investigate the effect on the miss rate for adults and children. In our work, we tried to augment the children's bounding boxes by applying horizontal flip.

Table 5 shows the detection performance when data augmentation is applied to the children's bounding boxes. It can be seen that when the batch size is 256, the miss rate for both attributes decreases. However, when the batch size is 64 or 128, the miss rate for children does not change, while the miss rate decreases for adults. These results indicate that applying data augmentation is effective in improving the overall detection performance. On the other hand, when we focus on the standard deviation, we must not forget that the disparity in detection performance between age attributes is expanding. First and foremost, a "person" can be an adult or a child. If the detection performance for adults is improved solely by increasing the data of children, we would consider that the bias in the quantity of data between classes is not directly relevant..

4.2 The relationship between the size of a person's bounding box and the miss rate

Detecting small objects is a difficult task in object and person detection research because of the limited information that can be obtained from a bounding box with a small image size. It is clear that children have smaller bodies than

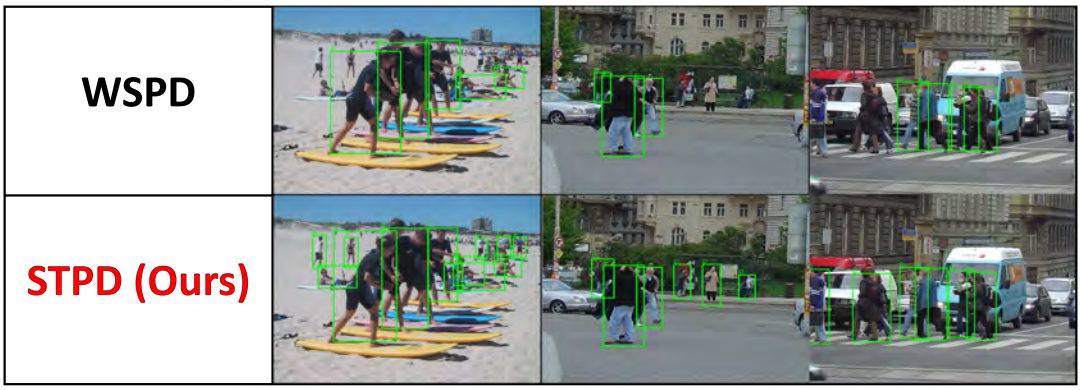

Figure 3: Comparison of detection results of WSPD and STPD.

Batch Size, Epochs	Miss Rate(Adult) [%]		Miss Rate(Children) [%]		Standard Deviation [%]	
	w/o Aug.	w/ Aug.	w/o Aug.	w/ Aug.	w/o Aug.	w/ Aug.
64, 10	13.8	**12.6**	19.2	19.2	**2.7**	3.3
128, 10	13.4	**12.1**	19.2	19.2	**2.9**	3.6
256, 10	13.1	**10.7**	17.3	**15.4**	**2.1**	2.4

Table 5: The impact of applying data augmentation (horizontal flip) only to the bounding boxes of the children in the training data. The results show that applying data augmentation is effective in improving the overall detection performance. On the other hand, it may increase the disparity in detection performance among age attributes.

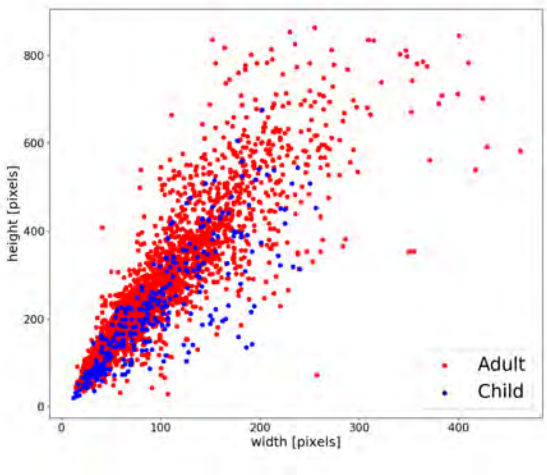

Figure 4: Distribution of bounding boxes for adults and children in the FA-INRIA Person Dataset. Children's bounding boxes tend to be relatively smaller than those of adults.

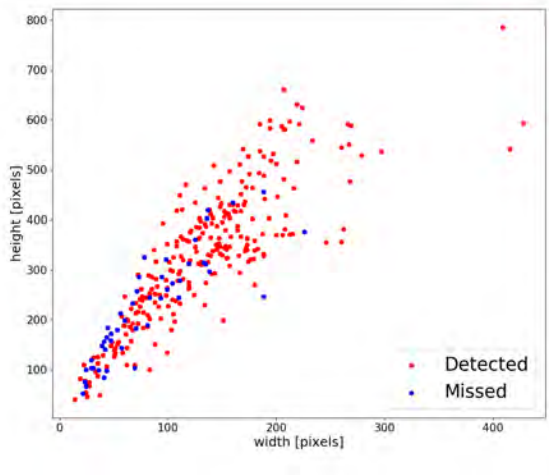

Figure 5: Whether bounding boxes can be detected in test data (red: detected, blue: missed).

adults. Therefore, the bounding boxes of children tend to be smaller than those of adults. Thus, we thought it would be important to investigate the size of the bounding boxes in the FA-INRIA.

Figure 4 presents the distribution of the size of the bounding boxes for adults and children. Adults are shown in red and children are shown in blue. This distribution indicates that most of the bounding boxes that exceed the size of 600 pixels * 300 pixels in height and width, respectively, are for adults. In other words, the difference in the size distribution of the bounding boxes may be one of the factors affecting the disparity in the miss rate. Figure 5 also shows the distribution of the size of the bounding boxes in the image for the FA-INRIA (test set): the bounding boxes that could be detected are shown in red, and the missed bounding boxes are shown in blue. As you can see in these figures, most of the missed bounding boxes are biased towards the smaller

Batch Size, Epochs	Miss Rate(Adult) [%]			Miss Rate(Children) [%]			Standard Deviation [%]		
	Input size of the image [pixels * pixels]								
	150	300	600	150	300	600	150	300	600
64, 10	14.9	**14.1**	14.4	17.3	**15.4**	**15.4**	1.2	0.7	**0.5**
128, 10	15.2	14.6	**13.9**	21.2	17.3	**15.4**	3.0	1.4	**0.8**
256, 10	14.1	14.7	**13.4**	21.2	17.3	**15.4**	3.6	1.3	**1.0**

Table 6: The effect of changing the input size of the image to the SSD on the detection performance for each age attribute. The results show that increasing the input size decreases the miss rate. In addition, children are more strongly affected by changes in the size of the input. We conclude that the bias in the size of the bounding box is a major factor in the disparity in detection performance.

image size. In other words, in order to mitigate the disparity in the miss rate further, it is necessary to use detectors that can detect small persons.

In this paper, we investigated the effect of changing the image size of the input on the miss rate of each attribute. The SSD resizes the input image to a set size, regardless of the size of the original image. This process is likely to result in missing details of the image. In order to detect small bounding boxes, we thought that increasing the size of the input image would suppress the missing information. We examined three patterns of input image sizes: (i) 150 pixels * 150 pixels, (ii) 300 pixels * 300 pixels, and (iii) 600 pixels * 600 pixels. The default size for the SSD is 300 pixels * 300 pixels. For more accurate validation, we also used a sub-dataset with the same number of bounding boxes for adults and children in the training data.

Table 6 shows the miss rate when the input image size of the SSD is changed. It can be seen that increasing the size of the input image is a major factor in reducing the miss rate. On the other hand, when the input size is small (150 pixels * 150 pixels), the miss rate for children is very poor. We consider that this is because image information is missing due to the relatively smaller bounding box as well. As shown in Figure 4, children's bounding boxes are more difficult to detect when the input size is small because children have a relatively higher proportion of small bounding boxes than adults. Based on this result and Figures 4 and 5, we conclude that the unbalanced distribution of the bounding box sizes is one of the main reasons for the disparity in detection performance between adults and children.

4.3 Appearance Difference

We have considered two aspects: the bias in the quantity of data between classes and the size of the bounding boxes. However, as shown in Figure 5, we can see that some people are not detected even though the bounding box is relatively large. Moreover, as mentioned in Section 4.1, we found that the bias in the quantity of data between classes is most likely not relevant. These results suggested that there might be other factors that generate disparities in detection performance between age attributes. Subsequently, we hypothesized that there would be apparent differences between the distributions of bounding box sizes of adults and children, as they differ significantly in size.

Figure 6 shows the compression of the image features using t-SNE and the visualization of the distribution. It is difficult to imagine that there is a disparity in detection performance based on the appearance of the distribution, which is not clearly divided by age attribute and is evenly distributed. This result supports the fact applying data augmentation to the children's bounding boxes was more effective in improving the detection rate for adults than for children. Since there is no apparent difference between adults and children, we reiterate that we do not need to consider the bias in the quantity of data between classes to reduce the miss rate for children.

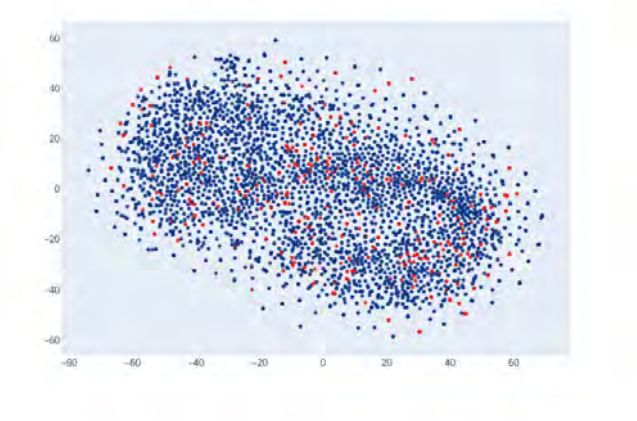

Figure 6: Data visualization of bounding boxes using t-SNE (blue: adults, red: children). There is no apparent significant difference between the bounding boxes of children and adults. As mentioned in Section 4.1, when data augmentation was applied to children's bounding boxes, the miss rate was strongly affected for adults but not for children. This data visualization supports the consideration that the bias in the quantity of data between classes has little to do with the disparity in detection performance.

5 Conclusion

In this paper, we investigated and examined various perspectives on the causes of the disparity in detection performance between adults and children in the task of pedestrian detection. As a first experiment, we confirmed that self-training extends the pre-training model and improves the overall detection performance. Next, we found that applying data augmentation to the bounding boxes of children, for whom there

is less data available than for adults, significantly improves the detection performance for adults but not children. We also visualized the feature distribution of the bounding boxes using t-SNE and found that was no apparent difference between adults and children. These results indicate that it is not necessary to consider the bias in the quantity of data in terms of age attributes in pedestrian detection.

On the other hand, when we looked at the size of the bounding boxes in our FA-INRIA, we observed that the distribution was biased toward a smaller size for children than for adults. In addition, we found that changing the input size of the image fed to the detector had a significant impact on the detection performance for children. In other words, we concluded that the disparity in the size of the bounding boxes was a major factor in the disparity in detection performance among the age attributes. In the future, focusing on the detection of small bounding boxes will help to mitigate the bias between attributes.

References

Brandao, M. 2019. Age and Gender Bias in Pedestrian Detection Algorithms. *arXiv preprint arXiv:1906.10490*.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 886–893. Ieee.

Girshick, R. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. SSD: Single Shot Multibox Detector. In *European conference on computer vision (ECCV)*, 21–37. Springer.

Minoguchi, M.; Okayama, K.; Satoh, Y.; and Kataoka, H. 2020. Weakly Supervised Dataset Collection for Robust Person Detection. In *arXiv pre-print:2003.12263*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Wilson, B.; Hoffman, J.; and Morgenstern, J. 2019. Predictive Inequity in Object Detection. *arXiv preprint arXiv:1902.11097*.

Zhao, Q.; Sheng, T.; Wang, Y.; Tang, Z.; Chen, Y.; Cai, L.; and Ling, H. 2019. M2det: A single-shot Object Detector Based on Multi-level Feature Pyramid Network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 9259–9266.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Zoph, B.; Ghiasi, G.; Lin, T.-Y.; Cui, Y.; Liu, H.; Cubuk, E. D.; and Le, Q. 2020. Rethinking Pre-training and Self-training. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 3833–3845. Curran Associates, Inc.