2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) | 978-1-7281-3079-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/FG47880.2020.00121

Noise Robust Hard Example Mining for Human Detection with Efficient Depth-Thermal Fusion

Zijian Zhao, Jie Zhang and Shiguang Shan
Key Lab of Intelligent Information Processing, Chinese Academy of Sciences (CAS)
Institute of Computing Technology, CAS, Beijing, China
{zijian.zhao, zhangjie, sgshan}@ict.ac.cn

Abstract— **Identity-preserving human detection is important for the privacy-protecting applications. IPHD[1] is a newly collected identity-preserving dataset that only contains depth and thermal images, which have much less information than RGB images. While less information and weakly labeled ground-truth boxes make it difficult to locate the objects correctly. In this paper, we adopt an efficient depth-thermal fusion approach to combine these two different inputs and enhance the representation. Moreover, a noise robust hard example mining algorithm is proposed to deal with weakly labeled data. The experiments show that our single model with single scale testing can get the AP=88.1 at IoU=0.5, which is a significant improvement compared with other competition results.**

I. INTRODUCTION

Human detection is a popular research topic of computer vision in recent years [2], [3], [4]. Human detectors also play important roles in many industrial applications such as ADAS, surveillance and other human involved visual tasks. There are many related benchmarks published. Most of them focus on the pedestrian detection, such as the classical works including INRIA [5], TudBrussels [6] and Daimler [7], as well as nowadays wildly used pedestrian dataset including Caltech-USA[8], KITTI[9] and CityPersons[10]. Besides that, there are also general human detection datasets proposed these years, e.g., CrowdHuman[11] and WiderPerson[12]. These two datasets have large amounts of images with high-density of people. Except for RGB image dataset, KAIST[13] provides thermal images together with color images for high accuracy pedestrian detection. While all of these human detection benchmarks call for color images, which is harmful to privacy protection. IPHD is a newly collected purely identity-preserving dataset that only contains depth images and thermal images[1]. It's worth exploiting privacy-protected human detection algorithms based on this dataset.

The thermal image and the depth image both contain less information compared with the color image, which will result in low accuracy for human detection. So it's important to make good use of the thermal and depth images together. Besides less information, this dataset is weakly labeled and exists plenty of inaccurate ground-truth boxes due to the thermal to depth registration, which makes it more difficult to learn a robust detector.

In this paper, we apply an efficient depth-thermal image early fusion approach which ensures sufficient cooperation between the depth and thermal information without increasing computational complexity. Moreover, we propose a noise robust hard example mining algorithm that make the learning process robust from the weakly labels, which is shown in Fig. 1. Equipped with these technologies, our model can achieve a significant improvement compared with the baselines and submitted results of the competition.

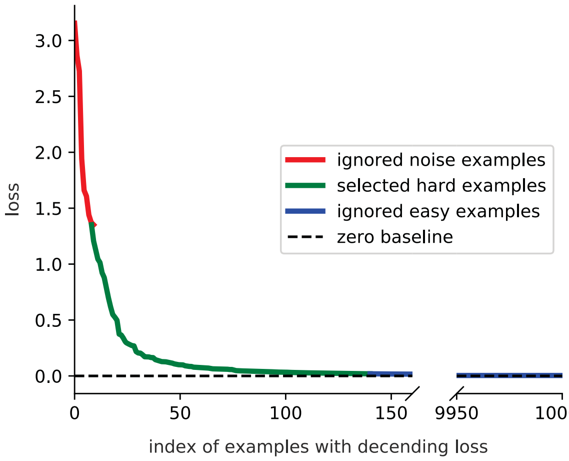


Fig. 1. Example of noise robust hard example selection approach. This curve is the descending loss values of top-10K negative proposals from one example image in a random selected training step. The red part is the N_{ign_n} ignored noise examples with large values of loss, the blue part is the ignored easy examples with small values of loss, and the green part represents the selected noise robust hard examples.

II. DATA DESCRIPTION AND PREPROCESSING

A. Data Description

In IPHD dataset, each pixel of the depth images contains absolute value of depth captured by RealSense D435. And the pixels of thermal images are temperature values captured by FLIR Lepton v3. The training set, validation set and testing set contain 84818, 12974 and 15115 pairs of depth and thermal frames respectively.

Weakly label. According to the official dataset description [1], annotation are done on the RGB images (not provided) captured by RealSense. Unlike depth images that are on-the-fly registered to RGB images by RealSense, thermal images are spatially registered to depth by human's off-line procedure, which inevitably introduces ground-truth boxes inaccuracy in thermal data.

978-1-7281-3079-8/20/$31.00 ©2020 IEEE
DOI 10.1109/FG47880.2020.00121

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on February 05,2021 at 04:12:54 UTC from IEEE Xplore. Restrictions apply.

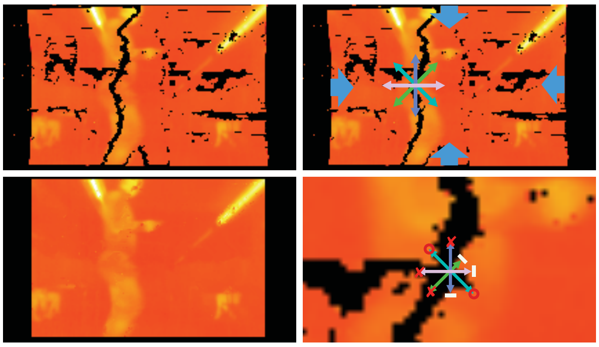


Fig. 2. Noise fixing method. The top-left: the original thermal image with noise. The bottom-left: the thermal image applied our noise fixing algorithm. The top-right: different directions of the kernel to deal zero values at the boundaries and inner image. The bottom-right: an illustration of inner filter with eight expanding directions in four pairs. The final fixed value is the mean of the center pixel values in red circles of the selected pair.

B. Data Normalization

Depth and Thermal images are different from the normal RGB or gray scale images. In order to initialize backbone with the ImageNet pre-trained parameters, the input data is truncated and normalized to make it similar to RGB one.

The pixel value of thermal images are rectified between 10 °C (283.15 K) and 40 °C (313.15 K) and multiply by $255/(31315 - 28315)$ to make it an unsigned integer. Environment temperature is estimated and subtracted from the image to get a sharp contrast between foreground and background. As for the depth images, the pixel values are rectified between 0.1 meters and 10 meters, and then they are also normalized to unsigned integer and subtracted by the mean value.

C. Noise Fixing

The thermal to depth registration brings many patches of zero value in the thermal images. Different kinds of kernels are applied to fill the zero patches with the reasonable values automatically. Fig.2 illustrates the visualization result and kernel examples of one image applied our noise fixing method.

III. Depth-Thermal Fusion Networks

A. Fusion Strategy

Depth images and thermal images have different distributions, so it is important to properly combine them. There are three practical fusion strategies for multiple input data, that is, early fusion, intermediate fusion and late fusion. Fig.3 shows how these three fusion strategies work.

In this task, early fusion, intermediate fusion and late fusion could be regarded as image level fusion, feature level fusion and box level fusion respectively.

Early fusion concatenates the thermal image and depth image along the channel dimension and treat the concatenated data as a new multi-channel image. In this work, we concatenate two duplicate thermal images and one corresponding depth image to create a 3-channel RGB-like image to make use of the ImageNet pre-trained weights for initialization.

Intermediate fusion has two separate feature extraction paths for thermal and depth information respectively. The features from these two paths are then merged by the feature fusion approach. Finally the merged feature is utilized as the intermediate part of the detection pipeline. Details of the feature fusion in our experiment can be found at the experiment section.

Late fusion applies two independent detectors that conduct the depth and thermal images detection separately. Then the detected boxes were merged by NMS. Besides NMS we used in this work, other boxes fusion methods for multi-model ensemble works could also be applied to merge the detected boxes from the two different branches.

Our experiments prove that early fusion has an advantage over the other two methods both in the final detection result and the computational complexity. In contrast to late fusion that only merge the detected boxes, early fusion make the depth and thermal information cooperate with each other during the feature extraction. This kind of cooperation empowers the model to extract and composite the useful information from depth and thermal inputs. It's worth noting that although intermediate fusion also conducts the merging procedure after the feature maps concatenation, while the merging of early fusion is achieved by a deep backbone network, which ensures more sufficient cooperation between the depth and thermal information.

B. Model Architecture

Fig.4 illustrates the model architecture that deploys early fusion strategy.

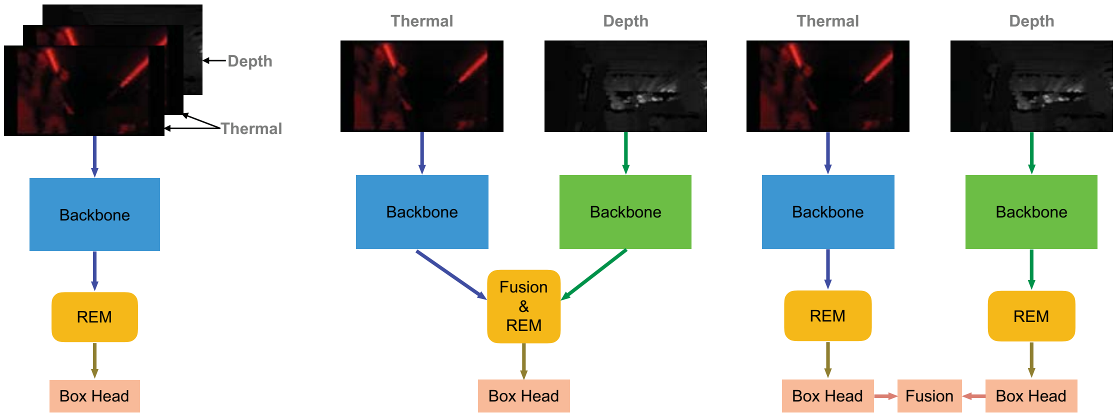
Backbone The model applies ResNet-50[15] as backbone to extract features from the depth and thermal input. There are five stages of residual blocks in original ResNet, that is, conv1, conv2, conv3, conv4 and conv5. Similar to FPN[15], the outputs from last residual blocks of conv3, conv4 and conv5 are denoted as {C3, C4, C5} respectively. These feature maps with stride {8, 16, 32} are employed as different levels to handle different object scales.

Receptive Enhancement Module (REM) FPN module is deprecated because both thermal and depth images only contain low level representation. To enhance the capability of multi-scale receptive fields without FPN, an inception module is applied to extract information with different spatial sizes. It has three inner branches with different kernel sizes of 3x3, 5x5 and 7x7, respectively.

Anchors There are three branches of detection heads with downsampling ratios {8, 16, 32} respectively. Multiple anchors with different scales {4xS, 6xS} and aspect ratios {0.5, 1, 2, 4} are designed for each branch, where S denotes the downsample ratio at a specific branch.

Box Prediction Module Two 1x1 sub-branches are attached on the inception module for classification and bounding box regression respectively.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on February 05,2021 at 04:12:54 UTC from IEEE Xplore. Restrictions apply.



(a) Early Fusion (b) Intermediate Fusion (c) Late Fusion

Fig. 3. Different fusion strategies of depth image and thermal image. REM reprecents Receptive Enhancement Module

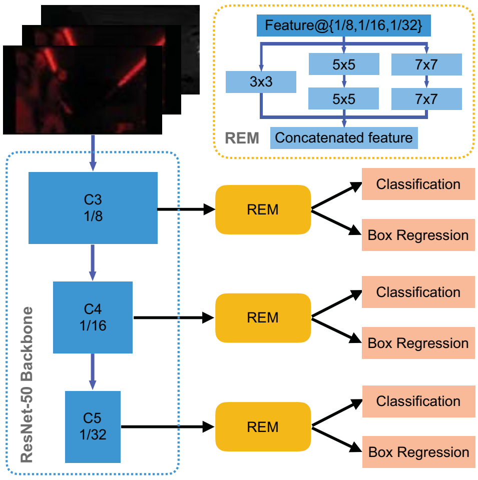


Fig. 4. Model architecture. The architecture of early fusion strategy and the detection pipeline with receptive enhancement module(REM).

C. Online Noise Robust Hard Example Mining

We adopt standard L1 loss for box regression[16] and 2-class softmax with cross-entropy loss for bounding box classification. The classification loss is defined as (1):

$$L_{cls} = \frac{1}{N_{pos}} \left[\sum_{i=N_{ign\_p}}^{N_{pos}} L_{cls\_p}^{sort}(\boldsymbol{p}_i, c_i) + \sum_{i=N_{ign\_n}}^{k*N_{pos}} L_{cls\_n}^{sort}(\boldsymbol{p}_i, c_i) \right], \tag{1}$$

where $L_{cls\_p}^{rank}$ is the descending sorted classification loss of positive samples and $L_{cls\_n}^{rank}$ is for the negatives.

Hard Examples Mining: There are more than 10 thousand anchors and most of them are negative ones. It introduces the imbalance between positive and negative examples for classification. Online hard example mining[17] is an effective method to deal with this imbalance. N_{pos} denotes the number of positive samples, k is the ratio between the negatives and positives for hard negative examples selection. In our experiments, we make k=7.

Noise Robust Filtering (NRF): The thermal images are weakly labeled caused by thermal to depth registration. So it is important to make the learning process robust to these weakly data. Unlike normal hard example mining approaches only focusing on the hard examples, our method ignores part of the top ranked losses, because the mislabeled or inaccurate boxes may result in high amplitude discrete losses. N_{ign_p} and N_{ign_n} in (1) represent numbers of positives and negatives that are considered to be noise examples. N_{ign_p} and N_{ign_n} are determined by the thresholds T_p and T_n for estimating noise examples. They are defined in (2):

$$\begin{aligned} T_p &= \frac{\lambda_p}{(\beta_p - \alpha_p) * N_{batch}} \sum\nolimits_{i=N_{batch}*\alpha_p}^{N_{batch}*\beta_p} L_{batch\_p}^{sort}[i], \\ T_n &= \frac{\lambda_n}{(\beta_n - \alpha_n) * N_{batch}} \sum\nolimits_{i=N_{batch}*\alpha_n}^{N_{batch}*\beta_n} L_{batch\_n}^{sort}[i], \end{aligned} \tag{2}$$

where N_{batch} denotes the batch size. $L_{batch\_n}^{sort}$ is the descending sorted loss of negative samples within a batch, and $L_{batch\_p}^{sort}$ is for the positive one. It's worth noting that $L_{batch\_\{p,n\}}^{sort}$ in (2) is different from $L_{cls\_\{p,n\}}^{sort}$ in (1): the former is batch level loss, while the latter is image level. Fig.1 shows an example of loss distribution curve of top-10k negative samples applied hard examples mining and noise robust filtering. $\lambda_{\{p,n\}}$ are scale factors for positives and negatives. $\alpha_{\{p,n\}}$ and $\beta_{\{p,n\}}$ are hyper-parameters that control the selected interval of sorted loss for noise threshold calculation. The relative appropriate values of these hyper-parameters in our experiments are set as $\{\lambda_p$=5, λ_n=3, α_p=3, β_p=5, α_n=6, β_n=10$\}$.

IV. Experiments

A. Training Details

Optimizer: We train the model using stochastic gradient descent (SGD) optimizer with momentum at 0.9 and weight decay at 5e-4 for totally 12 epochs. The learning rate starts from 1e-3 and divided by 10 at 7 and 9 epochs respectively. The model is trained on two NVIDIA TITAN Xp GPUs with 8 pairs of thermal and depth images on each card.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on February 05,2021 at 04:12:54 UTC from IEEE Xplore. Restrictions apply.

TABLE I

ABLATION ON DIFFERENT INPUT AND NOISE ROBUST FILTER.

#	D	T	NRF	AP@0.25	AP@0.5	AP@0.75
1	✓			81.4	75.2	44.0
2		✓		82.0	76.2	46.6
3		✓	✓	86.8	81.1	49.8
4	✓	✓	✓	**91.7**	**88.1**	**59.5**

TABLE II

EFFECT OF RECEPTIVE ENHANCEMENT MODULE.

REM	AP@0.25	AP@0.5	AP@0.75
	90.1	86.5	54.7
✓	**91.7**	**88.1**	**59.5**

Network Initialization: The ResNet-50 backbone is initialized with the parameters pre-trained on ImageNet. Except for the backbone, other modules such as REM module and box prediction module are trained from scratch.

Training Data The results are evaluated on the testing data, so we combine the original training data and validation data for model training.

Data Augmentation: Each image in a training batch applies the basic color distortion and random horizontal flip. Multiple scale training is applied to make up for the deficiency of scale variety. We randomly resize the short size of the images to [128, 256, 512] for each batch and keep the width-height ratio a constant value.

B. Inference Details

During testing phase, only one scale image is forwarded through the network. The image scale we choose is 256x448. Each 3-channel image is made up from two duplicate thermal images and one corresponding depth image. The output predicted boxes are then conducted deduplication by normal non-maximum suppression (NMS). Only single scale testing is applied to reduce the gap with industrial applications.

C. Ablation Study

Thermal, depth and fusion: In this section, we study the different performance of (a) only depth image as input, (b) only thermal image as input, and (c) depth and thermal fusion. Early fusion approach just treat the concatenated image as the 3-channel image. We make three copies of a depth image and concatenate them as a new 3-channel image when conducting the depth image human detection on depth images. The same thing is done for the thermal one. In Table.I, we denote depth as D, thermal as T and noise robust filtering as NRF. The result shows that the depth-thermal fusion method achieve significant performance improvement than only use depth or thermal images.

Effect of Noise Robust Filtering: Thermal data is weakly labeled caused by thermal to depth registration. The 2^{ed} and 3^{rd} row in Table.I show that noise robust filtering improves AP of thermal image detection by 4.8, 4.9 and 3.2 at three IOU thresholds respectively. Noise robust filtering is proved to be effective on the task with weakly labeled data.

TABLE III

COMPARISON OF DIFFERENT FUSION METHODS.

	AP@0.25	AP@0.5	AP@0.75
Early fusion	**91.7**	**88.1**	**59.5**
Intermediate fusion	88.5	83.5	51.4
Late fusion	90.1	84.0	47.6

TABLE IV

COMPARISON OF SUBMITTED COMPETITION RESULTS. RANK NO.3 , RANK NO.3 AND RANK NO.3 ARE TOP3 RESULTS OF THE COMPETITION.

	AP@0.25	AP@0.5	AP@0.75
iphd_baseline[19]	78.3	65.7	23.4
rank No.3[19]	79.5	71.8	40.2
rank No.2[19]	82.6	75.1	40.0
rank No.1[19]	84.3	81.7	54.7
ours	**91.7**	**88.1**	**59.5**

Receptive Enhancement Module: As is shown in Table.II, REM module improves AP by 0.4, 0.9 and 2.3 at IOU threshold equaling 0.25, 0.5 and 0.75 respectively. Based on the above, one can draw a conclusion that REM module improves the accuracy of bounding box localization.

D. Comparison of Different Fusion Strategies

We evaluated three fusion strategies shown in Fig.3, which are early fusion, intermediate fusion and late fusion. In the intermediate fusion approach, the model applies parallel ResNet-50 branches as backbones to extract features from depth and thermal images respectively. Features from these two ResNet50 backbones are concatenated together and then merged by a 3x3 convolutional layer. In the late fusion approach, the outputs from two independent detectors of depth and thermal images are merged by normal NMS.

Table.III shows the results of these three fusion strategies. As is analyzed before, early fusion shows significant advantages than the other two fusion strategies.

E. Comparison of Submitted results

Table.IV shows the comparison with the submitted results in the IPHD challenge[18], [19]. Our proposed single model with single testing scale achieves a significant improvement compared with the baseline and the top three final results.

V. CONCLUSIONS

In this work, we design a depth-thermal fusion model for identity-preserved human detection. The experiments are based on the IPHD dataset in ChaLearn Looking at People's competition organized in the context of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). An optimal fusion strategy is analyzed and the efficient detection pipeline is designed for depth-thermal human detection. To deal with the weakly labeled ground-truth boxes, an effective example selection method is proposed, which combines online noise example filtering and online hard example mining together. Our model with single-scale testing can gain significant improvement compared with the submitted results of the competition.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on February 05,2021 at 04:12:54 UTC from IEEE Xplore. Restrictions apply.

References

[1] Albert Claps, Julio C. S. Jacques Junior, Sergio Escalera and Carla Morral. ”Identity-preserved Human Detection (FG’20).” http://chalearnlap.cvc.uab.es/dataset/34/description/. 2020.

[2] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. ”Is faster r-cnn doing well for pedestrian detection?.” *In European conference on computer vision*, pp. 443-457. Springer, Cham, 2016.

[3] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi and In So Kweon, Multispectral Pedestrian Detection: Benchmark Dataset and Baseline, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[4] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. ”Repulsion loss: Detecting pedestrians in a crowd.” *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7774-7783. 2018.

[5] Navneet Dalal and Bill Triggs. ”Histograms of oriented gradients for human detection.” *In 2005 IEEE computer society conference on computer vision and pattern recognition*, vol. 1, pp. 886-893. IEEE, 2005.

[6] Christian Wojek, Stefan Walk, and Bernt Schiele. ”Multi-cue onboard pedestrian detection.” *In 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 794-801. IEEE, 2009.

[7] Markus Enzweiler, and Dariu M. Gavrila. ”Monocular pedestrian detection: Survey and experiments.” *IEEE transactions on pattern analysis and machine intelligence* 31, no. 12 (2008): 2179-2195.

[8] Piotr Dollr, Christian Wojek, Bernt Schiele, and Pietro Perona. ”Pedestrian detection: A benchmark.” *In 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 304-311. IEEE, 2009.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. ”Are we ready for autonomous driving? the kitti vision benchmark suite.” *In 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354-3361. IEEE, 2012.

[10] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. ”Citypersons: A diverse dataset for pedestrian detection.” *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213-3221. 2017.

[11] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. ”Crowdhuman: A benchmark for detecting human in a crowd.” *arXiv preprint* arXiv:1805.00123 (2018).

[12] Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z. Li, and Guodong Guo. ”WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild.” *IEEE Transactions on Multimedia (2019)*.

[13] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. ”Multispectral pedestrian detection: Benchmark dataset and baseline.” *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1037-1045. 2015.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. ”Deep residual learning for image recognition.” *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[15] Tsung-Yi Lin, Piotr Dollr, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. ”Feature pyramid networks for object detection.” *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125. 2017.

[16] Ross Girshick. ”Fast r-cnn.” In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.

[17] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. ”Training region-based object detectors with online hard example mining.” *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761-769. 2016.

[18] Albert Claps, Julio C. S. Jacques Junior, Sergio Escalera and Carla Morral. ”2020 Looking at People Challenge FG Identity-preserved human detection.” http://chalearnlap.cvc.uab.es/challenge/34/description/. 2020.

[19] Albert Claps, Julio C. S. Jacques Junior, Sergio Escalera and Carla Morral. https://competitions.codalab.org/competitions/21928#results. 2020.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on February 05,2021 at 04:12:54 UTC from IEEE Xplore. Restrictions apply.