HeadNet: Pedestrian Head Detection Utilizing Body in Context

Gang Chen[1,2], Xufen Cai[1], Hu Han[*,1], Shiguang Shan[1,2,3] and Xilin Chen[1,2]
[1] *Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China*
[2] *University of Chinese Academy of Sciences, Beijing 100049, China*
[3] *CAS Center for Excellence in Brain Science and Intelligence Technology*
{*gang.chen, xufen.cai*}*@vipl.ict.ac.cn,* {*hanhu, sgshan, xlchen*}*@ict.ac.cn*

***Abstract*—Pedestrian head with arbitrary poses and size is prohibitively difficult to detect in many real world applications. An appealing alternative is to utilize object detection technologies, which tend to be more and more mature and faster. However, general object detection technologies can hardly work in complicated scenarios where many heads are often too small to detect. In this paper, we present a novel approach that learns a semantic connection between pedestrian head and other body parts for head detection. Specifically, the proposed model, named as HeadNet, is based on PVANet backbone and also introduces beneficial strategies including online hard example mining (OHEM), fine-grained feature maps, RoI Align and Body in Context (BiC). Experiments demonstrate that our approach is able to utilize spatial semantics of the entire body effectively, and gains inspiring performance for pedestrian head detection.**

I. Introduction

Pedestrian head detection plays an import role in security applications such as pedestrian detection or people counting, etc. Current research mainly focuses on the detection of the whole pedestrian, without taking spatial relations of individual body parts into account. In this paper, we propose an approach to detect pedestrian head along with pedestrian, head-shoulder, and upper body by utilizing body in context. The spatial relations between head and other body parts will have significant effect. Pedestrian head detection using spatial semantic relations is more challenging in the following aspects: Firstly, pedestrian head in complex scenarios is usually too small to detect and a number of heads may present, e.g., pedestrian head detection at distance for the crowded scene or its application in fast people counting. Secondly, a head is a part of pedestrian and there is not known methods to learn their spatial correlation for better head detection performance. Also, the scarcity for this kind of datasets is another makes this problem more difficult.

While great progress has been made on object detection recently, the detection of pedestrian head is still in its infancy. Most previous works of object detection are constrained to the detection of people or pedestrian and limited studies focusing on pedestrian head detection have been reported. Deep learning methods of object detection can be mainly divided into two categories. One is based on region proposal. Since 2014, Girshick et al. have proposed a CNN and region proposal based framework name R-CNN [1] replacing the traditional detection method at the first time, which has made a great breakthrough in object detection. R-CNN first extracted about 2,000 region proposal by selective search [2], and then scaled the image cropped by each region proposal into 227×227 before as the input of CNN. Finally, the output of layer fc7 was entered into SVM for classification. In this way, R-CNN significantly improved mAP from 34.3% (DPM HSC [3] [4] [5]) to 66% in PASCAL VOC 2007 [6]. It's worth nothing that each region proposal in R-CNN requires calculating deep feature one time. For this reason, He et al. and Girshick et al. have proposed SPP-NET [7] and Fast R-CNN [8], respectively. A image can be extracted only once, and the corresponding features are extracted according to the position of region proposals in feature maps, which made the detection speed 160 times faster than that of R-CNN. Followed by that, Ren et al. have proposed regional proposals network (RPN) in Faster R-CNN [9] to generate regional proposals of high quality rather than selective search. RPN and Fast R-CNN are shared the backbone network and combined into Faster R-CNN. Faster R-CNN not only improved the detection accuracy, but also accelerating the detection speed of each frame from 2s to 192m. Dai et al. have proposed R-FCN [10] by merging Faster R-CNN and Residual Net [11] motivated by its remarkable performance in image classification. Kong et al. have proposed HyperNet [12] by fusing the feature maps of multiple layers as hyper feature fed to RPN and classification network, which in turn improved the precision of object detection. After that, Kim et al. have proposed PVANet [13] based on the principle of more layers and less channels. It not only utilized the residual substructures of Inception [14] and C.Relu but also introducing the multi-scale representing of hyper feature, PVANet achieved competitive accuracy compared with Faster R-CNN, but processing speed did not exceed 50 ms per frame.

Another line of research for object detection is based

*H. Han is the corresponding author.

978-1-5386-2335-0/18/$31.00

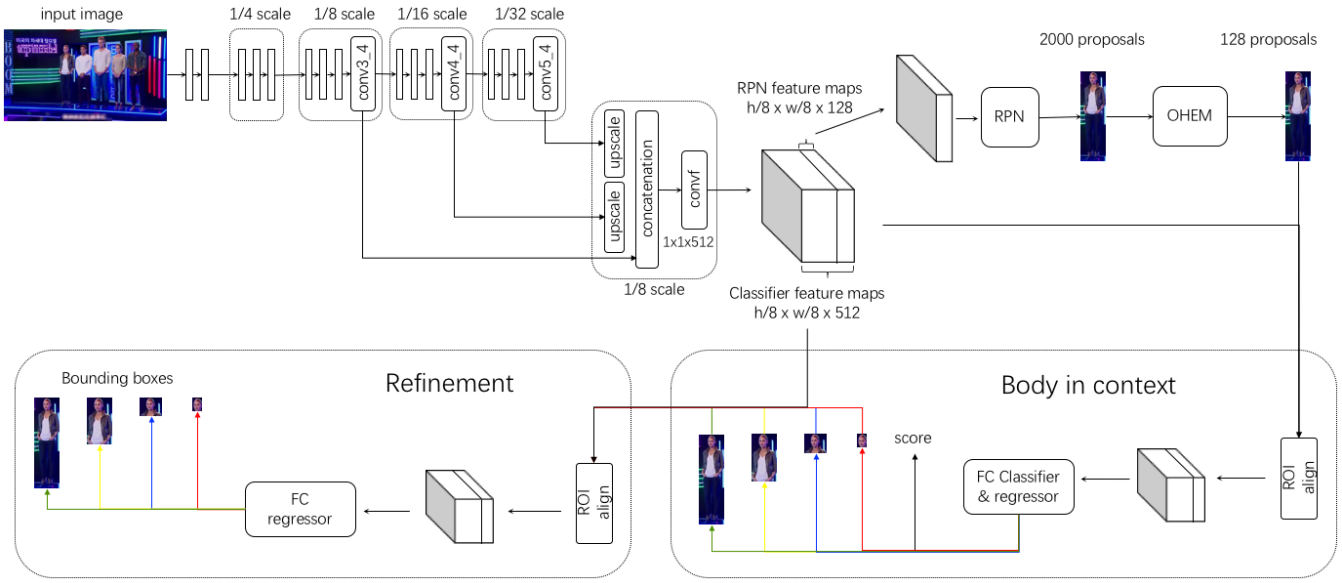


Figure 1: The architecture of the proposed HeadNet for pedestrian head detection in unconstrained scenario.

on regression. Redmon et al. have proposed YOLO [15] to generate a feature map in size of 7×7, which enabled each grid on the grid on the feature map used to regress the target location and category. At the same time, Liu et al. have also proposed SSD [16] to regress the target location and category, but utilize multi-scale feature maps.

In this paper, we aim to detect small and hard-to-detect pedestrian head using spatial semantic relations along with another three related strategies for object detection. This paper is the first research attempt to jointly regress head, head-shoulder, and upper body through pedestrian with more fine-grained features, thereby resulting in more reliable head detection.

II. Our Approach

For hard-to-detect pedestrian head detection problem, we propose HeadNet to learn the semantic relations among head and other body parts in context. Our HeadNet introduces beneficial OHEM [17], fine-grained feature maps, RoI Align strategies and Body in Context (BiC) to achieve robust head detection. The specific framework is given in Fig. 1 and Subsection A. Subsection B provides the insights of corresponding strategies in detail, and Subsection C elaborates the loss function of HeadNet.

A. Framework

Our HeadNet framework is illustrated in Fig. 1. For a input image fed to HeadNet, we first extract 1/8 size feature maps by multi-scale feature fusion of the deep convolution layers. Based on these feature maps, RPN network will be utilized to generate 2000 pedestrian proposals. OHEM is introduced to sample 128 hard proposals. Then we design Body-in-context (BiC) part for joint detection of head and other body parts from coarse to fine. More specifically, we use RoI Align [18] to extract the corresponding features of 128 hard proposals followed by simultaneously predicting the probability of proposals to be positive and regressing bounding boxes of head, head-shoulder, upper body and pedestrian. Last but not least, we design the Refinement part to achieve a more accurate detection for pedestrian head. Refinement part also takes advantage of joint predictions of multi-components, which further regresses bounding boxes of each components. The network parameters of BiC and Refinement part are shared with each other.

Overall, in HeadNet, we introduce OHEM to find hard pedestrian proposals from extremely unbalanced proposal distribution. Aiming at the problem that many pedestrian heads are too small to detect, we design fine-grained feature maps to fit the size of pedestrian head. In this way, we can obtain feature maps with sufficient information for pedestrian heads that are rather small. Moreover, RoI Align is utilized in both BiC and Refinement parts instead of RoI Pooling to sample multi-scale feature from fine-grained feature maps.

B. Backbone

In HeadNet, we use PVANet [13] as backbone. Kim et al. proposed this more lightweight framework that has the same precision as other state-of-the-art networks. This new deep neural network reduces redundancy by adopting the latest structures of C.ReLU and Inception [14]. It scored 84.9% and 84.2% respectively on PASCAL VOC 2007 and PASCAL VOC 2012, while the calculation cost is 10% less than that of ResNet-101.

C. Strategies in Details

1) Online hard example mining: [17] introduces OHEM used in Fast R-CNN, which selects hard RoIs by calculating the sum of both classification and regression loss in advance, rather than randomly selecting according to the rate $1:3$ of positive and negative samples. In HeadNet, we also detect based on region proposal, first produce hundreds of thousands of proposals by RPN network, and then select the first 12000 proposals before non-maximum suppression (NMS) followed by picking up the top 2000 proposals according to probability values. Finally, the 128 proposals with the largest loss are selected as the training samples of the network, which makes the network training have more attention to a small number of difficult samples in a large number of simple background samples, thus making the network more effective.

2) Fine-grained feature maps: We upscale the output of layers conv4_4 (1/16 size of original image) 2 times and conv5_4 (1/32 size of original image) 4 times, which have abundant semantic information, into 1/8 size of original image by deconvolution filter. Instead of downsampling conv3_4 layers in PVANet, we remain the size of conv3_4 unchanged to keep their abundant spatial information. Then the output of layers conv3_4, conv4_4 and conv5_4 are combined together to get multi-scale fused feature maps that are sophisticated in both semantics and space.

3) RoI Align: [18] introduced RoI Align which is an improvement of RoI Pooling. The input size of fully connection layer is fixed (e.g., 7×7), while each the size of the RoI is arbitrary. RoI Pooling maps the RoI in the original image into feature maps and extract the features with fixed size. For feature maps are 1/8 of the original size and the features are discrete in maps, RoI Pooling have to scale the original 1/8 of RoI into a float type then rounding the scaled RoI region to achieve relatively coarse sampling in space. In order to get more sophisticated feature, RoI Align calculates the feature's value on the float coordinate by bilinear interpolation rather than rounding the scaled RoI region.

4) Body in Context (BiC): In this part, we introduce BiC part into HeadNet, which attempts to detect pedestrian head from coarse to fine. Firstly, pedestrian's RoI features are extracted from the feature maps using pedestrian regional proposals by RoI Align. These features are used to calculate the probability of the proposals as part of the pedestrian, and regress the head, head-shoulder, upper body and pedestrian of the proposals. Secondly, in the Refinement part accompany with the all proposals produced by BiC and sample features using RoI Align again, the more sophisticated RoI features are extracted to regress the proposals of head, head-shoulder, upper body and pedestrian in depth.

D. HeadNet Loss

Multi-task learning has been proved to be effective in learning sharable and robust feature in a number of computer vision tasks [19]–[21]. The loss of HeadNet consists of multi-tasks losses, including RPN loss, BiC loss and Refinement loss listed as (1):

$$L_{HeadNet} = L_{RPN} + \lambda_1 L_{BiC} + \lambda_2 L_{Refine}, \tag{1}$$

For RPN loss, we adopt the original loss function in [9].

BiC loss is made of two parts, one part is the classification loss of pedestrian proposals and the other part is the regression loss from pedestrian proposals to pedestrian, head, head-shoulder, and upper body proposals listed as (2):

$$\begin{aligned} L_{BiC} &= \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) \\ &+ \lambda \sum_{j \in \{parts\}} \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_{i,j}, t_{i,j}^*), \end{aligned} \tag{2}$$

where $L_{cls}(p_i, p_i^*)$ is classification loss and $L_{reg}(t_{i,j}, t_{i,j}^*)$ is regression loss. Classification loss for BiC is formulated as (3):

$$L_{cls}(p_i, p_i^*) = -p_i^* \log(p_i) - (1 - p_i^*) \log(1 - p_i), \tag{3}$$

where for the i^{th} pedestrian proposals, p_i denotes the prediction probability of positive pedestrian proposals, p_i^* denotes the ground truth of pedestrian proposal, which $p_i^* = 1$ denotes the proposals are positive otherwise negative.

Regression loss for BiC is formulated as (4):

$$L_{reg}(t_{i,j}, t_{i,j}^*) = \sum_{k \in \{x,y,w,h\}} \text{smooth}_{L_1}(t_{i,j,k} - t_{i,j,k}^*), \tag{4}$$

where for the i^{th} pedestrian proposals, $t_{i,j,x}^*, t_{i,j,y}^*, t_{i,j,w}^*, t_{i,j,h}^*$ represent the offset of the real value from the i^{th} pedestrian proposals to the j^{th} part of pedestrian, head, head-shoulder, and upper body proposals respectively and are calculated by the following (5-8):

$$t_{i,j,x}^* = (x_{i,j}^* - x_{i,a})/w_{i,a} \tag{5}$$

$$t_{i,j,y}^* = (y_{i,j}^* - y_{i,a})/h_{i,a} \tag{6}$$

$$t_{i,j,w}^* = \log(w_{i,j}^*/w_{i,a}) \tag{7}$$

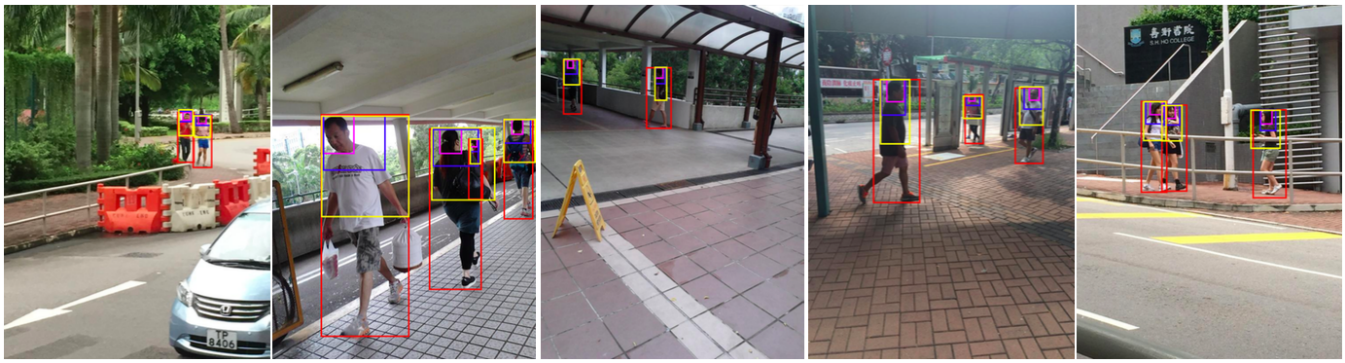
$$t_{i,j,h}^* = \log(h_{i,j}^*/h_{i,a}) \tag{8}$$

here we have

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}. \tag{9}$$

For Refinement loss, it only contains regression loss listed as (10), which represents the regression errors from the outputs of BiC network to the corresponding real positions of pedestrian, head, head-shoulder, and upper body.

(a)



(b)

Figure 2: Some example images from CUHK-SYSU Person Search Dataset. Pink, blue, yellow and red bounding boxes stand for the ground truth annotations of head, head-shoulder, upper body and pedestrian, respectively. (a) Examples of original annotation dataset; (b) Examples of the complementary annotations we provided.

Table I: Experiments setting

Network	Backbone	Pre-train with ImageNet	Proposals	using OHEM	Multi-scale training
Faster R-CNN	VGG-16	yes	300	no	no
R-FCN-50	ResNet-50	yes	300	yes	no
R-FCN-101	ResNet-101	yes	300	yes	no
PVANet	PVANet	yes	200	no	yes
HeadNet	PVANet	yes	200	yes	yes

$$L_{Refine} = \sum_{j\in\{parts\}} \frac{1}{N_{reg}} \sum_{i} p_i^* L_{reg}(t'_{i,j}, t'^*_{i,j}), \qquad (10)$$

III. Experiments

A. Datasets Description

The experimental dataset is originally from CUHK-SYSU Person Search Dataset [22], which consists of 18,184 images collected from streets and films with pedestrian bounding box annotations alone. We annotate 16,907 images among this database with head, head-shoulder, and upper body bounding boxes, and divide it into the training set and test set, which contain 13399 and 3508 images, respectively.[1] We randomly sample some images from the dataset and show them in Fig. 2. We can see the significant variations in the perspective, illumination and background in these images.

[1]The dataset with the annotations we provided is available at https://github.com/Dataset-VIPL-CAS/PSDBC

B. Implementation Details

We initialize HeadNet using the parameters of PVANet which pre-trained on ImageNet [23] and the fine-tune HeadNet on the renewed CUHK-SYSU Person Search Dataset training set. For better detection performance of different sizes of images and pedestrian head, we adopt multi-scale training strategy. When training, the shorter edge of images is scaled into between 416 pixels and 864 pixels while the longer edge does not exceed 1440 pixels. We solve the objective by SGD method, with the learning rate set to 0.001 for the beginning 50,000 runs and 0.0001 for the later 50,000

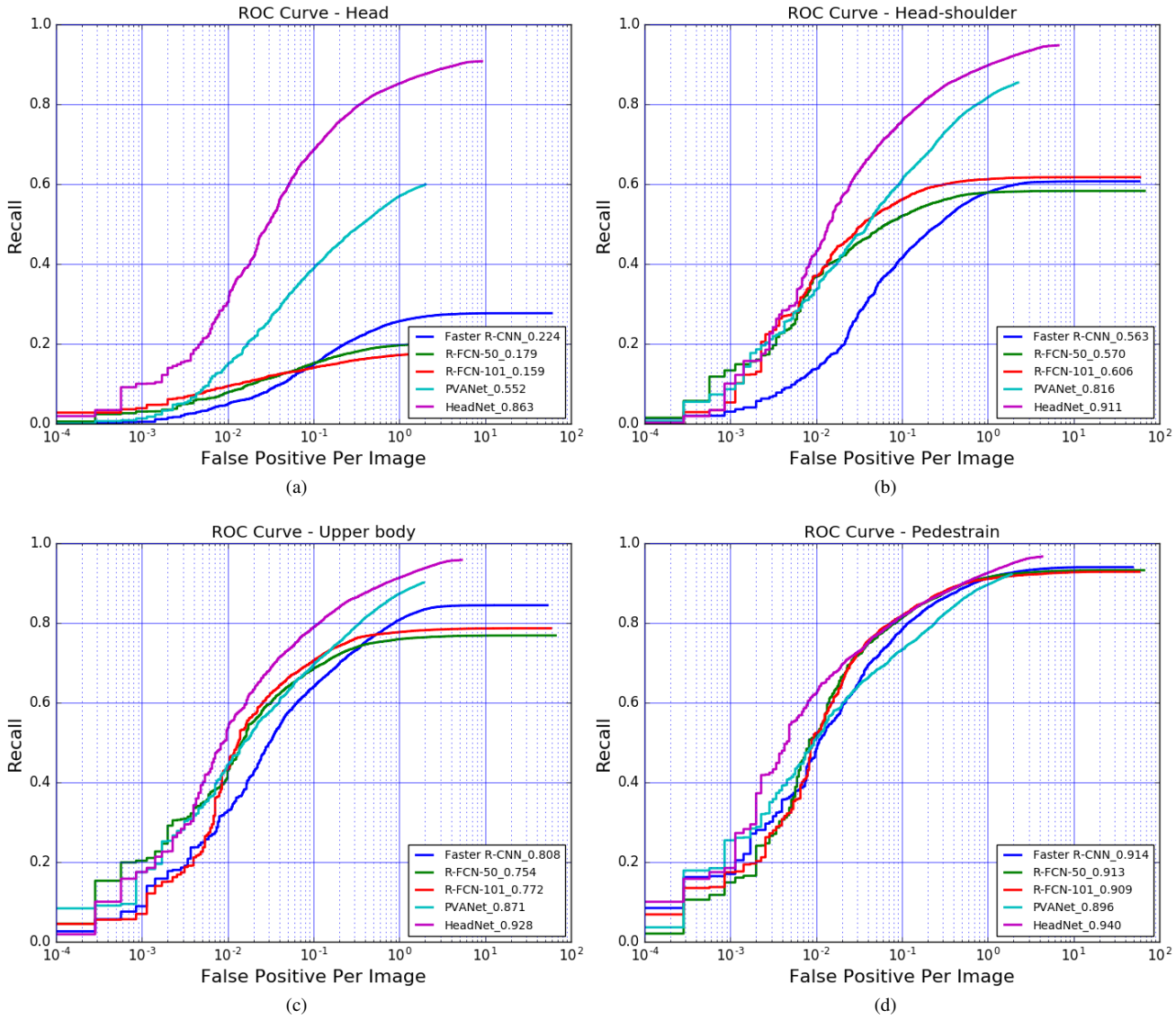


Figure 3: The detection results on the CUHK-SYSU Person Search Dataset. (a) Head's ROC curve, (b) Head-shoulder's ROC curve, (c) Upper body's ROC curve, and (d) Pedestrian's ROC curve.

runs. The hyper parameters of loss function, i.e., $\lambda_1, \lambda_2, \lambda$, are set to 0.25, 0.25 and 1, respectively.

For model comparisons, we train Faster R-CNN [9], R-FCN-50 [10], R-FCN-101 [10] and PVANet [13] use the same Caffe environment. The settings for all five networks are listed in TABLE I. We adopt VGG-16 [24], ResNet-50 [11] and ResNet-101 [11] as the backbone of Faster R-CNN, R-FCN-50 and R-FCN-101, respectively. All models are pre-trained on ImageNet [23]. In the stage of training, we take head, head-shoulder, upper body and pedestrian as four classes of bounding boxes as the input for Faster R-CNN, R-FCN-50, R-FCN-101 and PVANet by end-to-end training. Faster R-CNN, R-FCN-50 and R-FCN-101 use single-scale training method, where the shorter edge of images is scaled into 600 pixels and the longer edge does not exceed 1000 pixels. PVANet use multi-scale training method the same as HeadNet. In the stage of testing, we test all five models in single-scale. Considering efficiency and accuracy, the HeadNet and PVANet only use the first 200 proposals given by RPN.

For strategy verification, we train another three models, namely PVANet-OHEM (PVANet with OHEM), PVANet-OHEM-Refined (PVANet with OHEM and fine-grained feature maps) and PVANet-OHEM-Refined-RoiAlign (PVANet with OHEM, fine-grained feature maps and Roi Align), which compare with previously trained PVANet and HeadNet (PVANet with OHEM, fine-grained feature maps, Roi Align, BiC and Refinement) to verify the effectiveness of

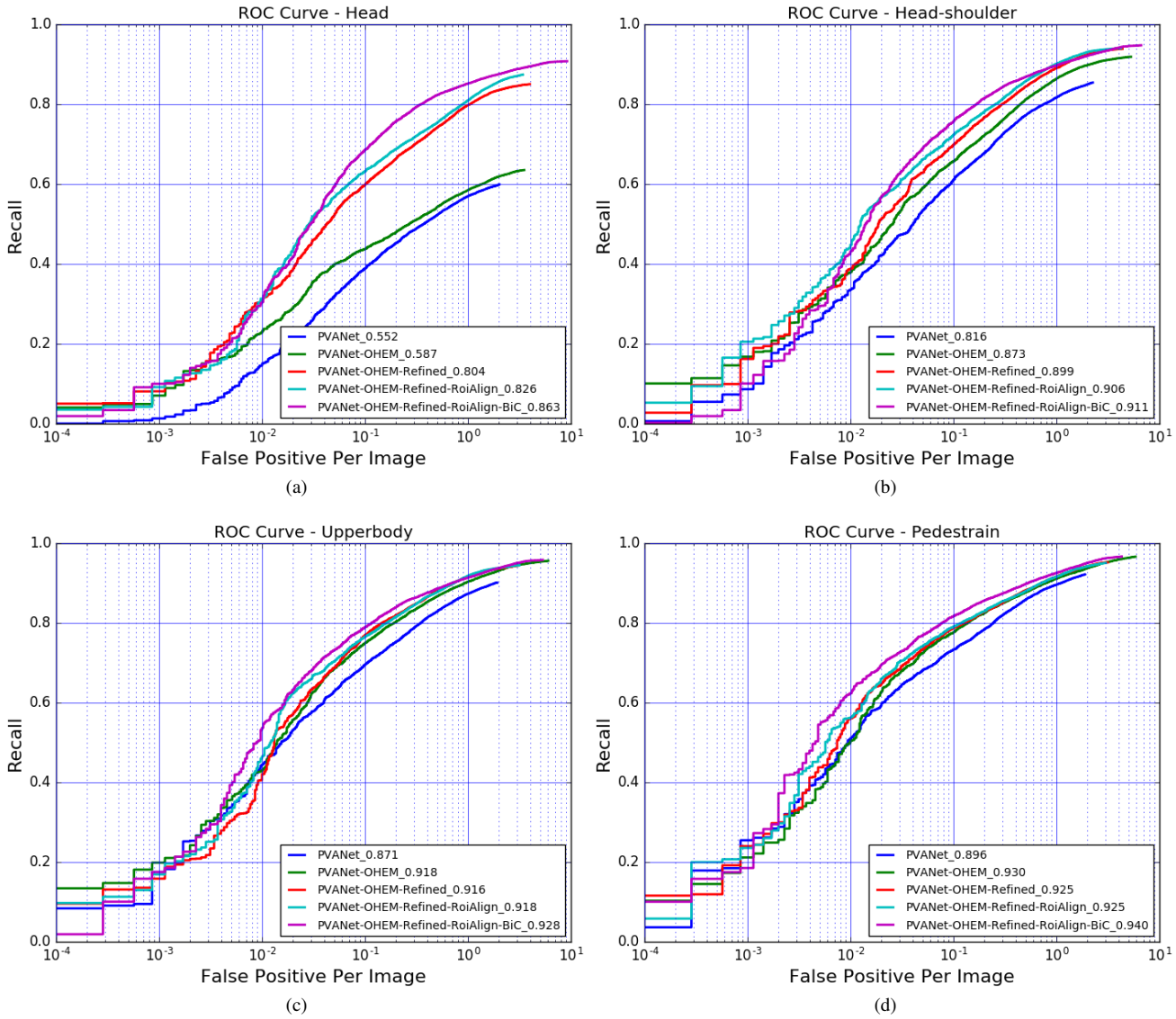


Figure 4: Strategy verification experiments on the CUHK-SYSU Person Search Dataset with and without using OHEM, refinement, and RoI Align in the proposed approach. (a) Head's ROC curve, (b) Head-shoulder's ROC curve, (c) Upper body's ROC curve, (d) Pedestrian's ROC curve.

our introduced strategies.

C. Model Comparisons

In this section, we compare HeadNet with Faster R-CNN, R-FCN-50, R-FCN-101 and PVANet in the testing set, the performance of all models are illustrated in Fig. 3. We quantitatively compare HeadNet to the state of the art. We observe from Fig. 3 that HeadNet outperforms PVANet by 31% increased from 55.2%, indicating that it is beneficial to utilize spatial semantics between head and body in context for head detection. Similarly, HeadNet also outperforms other models for head-shoulder detection, upper body detection and pedestrian detection. This observation indicates that different related body parts should be utilized jointly. We also observe that Faster R-CNN, R-FCN-50 and R-FCN-101 do not work very well for head detection although they perform fairly well for pedestrian detection as shown in Figs. 3 (a) and (d). This experiment demonstrates that our model is robust for all body parts and the whole pedestrian.

D. Strategies Verification

In this section, we analyze the impact of individual strategies in HeadNet, namely OHEM, fine-grained feature maps, RoI Align and BiC. We add these four strategies into PVANet gradually to get the following models: PVANet-

Table II: Comparisons of different proposal sampling methods.

	Pedestrian AP	Upper body AP	Head-shoulder AP	**Head AP**
Heuristic	89.6	87.1	81.6	**55.2**
OHEM	93.0	91.8	87.3	**58.7**
Gain (%)	+3.4	+4.7	+5.7	**+3.5**

Table III: Comparisons of using and not using fine-grained features.

	Pedestrian AP	Upper body AP	Head-shoulder AP	**Head AP**
Original	93.0	91.8	87.3	**58.7**
Fine-grained	92.5	91.6	89.9	**80.4**
Gain (%)	-0.5	-0.2	+2.6	**+21.7**

OHEM, PVANet-OHEM-Refined, PVANet-OHEM-Refined-RoiAlign, PVANet-OHEM-Refined-RoiAlign and HeadNet. The ROC curves for the head, head-shoulder, upper body and pedestrian are shown in Fig. 4. Note that models are compared progressively, which demonstrates the effectiveness of different strategies. As can be observed from the performance of head, head-shoulder, upper body and pedestrian, the model using all strategies always achieved better performance than other models. Fig. 4 also exhibits a clear trend that when the new strategy is added, the advantage becomes more striking, especially for fine-grained feature maps of head detection illustrated in Fig. 4 (a).

1) OHEM: TABLE II lists the performance after adding OHEM at the stage of proposals sampling. For original proposals generated by RPN, the distribution of samples is extremely imbalanced. The number of negative examples is far greater than that of the positive examples. Moreover, there are containing a large number of simple samples. Heuristic based proposals sampling method focuses on tackle the imbalanced problem by random sampling from both negative and positive proposals while OHEM based method exploit the hard examples by calculating the loss, therefore it avoid the gradients of hard examples are covered by a large number of simple samples in the training stage. Results show that OHEM improves the overall performance a lot.

Table IV: Comparisons of different feature pooling methods.

	Pedestrian AP	Upper body AP	Head-shoulder AP	**Head AP**
Roi Pooling	92.5	91.6	89.9	**80.4**
Roi Align	92.5	91.8	90.6	**82.6**
Gain (%)	+0.0	+0.2	+0.7	**+2.2**

Table V: Comparisons between using and not using BiC.

	Pedestrian AP	Upper body AP	Head-shoulder AP	**Head AP**
Original	92.5	91.8	90.6	**82.6**
BiC	94.0	92.8	91.1	**86.3**
Gain (%)	+1.5	+1.0	+0.5	**+3.7**

2) Fine-grained feature maps: TABLE III lists the performance after using fine-grained feature maps. In order to solve the key problem that pedestrian head is too small to detect, we cancelled the down-scaling process and design more fine-grained features during feature fusion. In this way, we can get features with higher resolution which results in more robust detection for extremely small targets. Results show that its performance on head detection has a significant improvement as high as 21.7%.

3) RoI Align: TABLE IV lists the performance after introducing RoI Align. Results show that it also greatly boost the performance of head detection. RoI Pooling may extract the mismatching regions or even loss more information, because small targets have rather small region on the feature maps, while RoI Align calculates more accurate feature values by bilinear interpolation.

4) Body in Context: TABLE V lists the performance after taking spatial semantic relations into account. Using body in context, we design a global to local detection method, which is detecting from pedestrian proposal to head, head-shoulder and upper body. Results show that the strategy not only achieves excellent performance for small targets such as head and head-shoulder, but also jointly promoting upper body and pedestrian detection.

IV. Conclusion

In this paper we have focused on how to utilize spatial semantic relations for detecting pedestrian head. The main challenge confronted is that most of the pedestrian heads are extremely small and the scenarios are also very complex. We propose to learn the spatial semantic relationship from coarse to fine and regress head and other parts jointly from the pedestrian proposal. Extensive experiments demonstrate that: 1) Fine-grained feature map resolution is more suitable for small objects, i.e., the hard-to-detect head; 2) Our model is able to effectively leverage the information from body in context for head detection.

In the future we would like to investigate graph models to model the semantic relationship of individual body parts, and apply the proposed HeadNet to people counting with arbitrary occlusions. We will also utilize more intra-part relationships of the pedestrian and the head, e.g., the symmetry property of human body and a body part to handle partial occlusions. In addition, utilizing 3D prior to improve the feature learning robustness from 2D images [25] will also be considered.

ACKNOWLEDGMENT

This research was supported in part by the National Basic Research Program of China (grant 2015CB351802), Natural Science Foundation of China (grants 61390511, 61732004, 61672496, 61650202, and 61702486), External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843), Strategic Priority Research Program of CAS (grant XDB02070004), and Youth Innovation Promotion Association CAS (grant 2018135).

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[2] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[4] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008, pp. 1–8.

[5] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester, "Discriminatively trained deformable part models, release 5," http://people.cs.uchicago.edu/˜rbg/latent-release5/, 2012.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *ECCV*, 2014, pp. 346–361.

[8] R. Girshick, "Fast R-CNN," in *ICCV*, 2015, pp. 1440–1448.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[10] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *NIPS*, 2016, pp. 379–387.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[12] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *CVPR*, 2016, pp. 845–853.

[13] S. Hong, B. Roh, K.-H. Kim, Y. Cheon, and M. Park, "PVANet: Lightweight deep neural networks for real-time object detection," *arXiv preprint arXiv:1611.08588*, 2016.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016, pp. 21–37.

[17] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *CVPR*, 2016, pp. 761–769.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017.

[19] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE TPAMI*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[20] F. Wang, H. Han, S. Shan, and X. Chen, "Deep multi-task learning for joint prediction of heterogeneous face attributes," in *FG*, 2017, pp. 173–179.

[21] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE TPAMI*, 2018.

[22] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *CVPR*, 2017.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2d face recognition via discriminative face depth estimation," in *ICB*, 2018, pp. 1–8.