Association Loss for Visual Object Detection

Dongli Xu, Jian Guan, *Member, IEEE*, Pengming Feng, *Member, IEEE*, and Wenwu Wang, *Senior Member, IEEE*

***Abstract*—Convolutional neural network (CNN) is a popular choice for visual object detection where two sub-nets are often used to achieve object classification and localization separately. However, the intrinsic relation between the localization and classification sub-nets was not exploited explicitly for object detection. In this letter, we propose a novel association loss, namely, the proxy squared error (PSE) loss, to entangle the two sub-nets, thus use the dependency between the classification and localization scores obtained from these two sub-nets to improve the detection performance. We evaluate our proposed loss on the MS-COCO dataset and compare it with the loss in a recent baseline, i.e. the fully convolutional one-stage (FCOS) detector. The results show that our method can improve the AP from 33.8 to 35.4 and AP_{75} from 35.4 to 37.8, as compared with the FCOS baseline.**

***Index Terms*—Association loss, object detection, object localization, object classification, convolutional neural networks.**

I. Introduction

VISUAL object detection aims to localize targets of interest and identify their categories from an image, which is required in a number of applications, such as face detection [1] and instance segmentation [2]. It often involves two tasks: object classification, i.e. to determine the class category of the object, and object localization, i.e. to determine its position in the image (e.g. the four coordinates of the bounding box centred at the object position), respectively.

The state of the art methods for this problem are based mostly on convolutional neural networks (CNN) such as the Faster R-CNN [3], YOLO [4], SSD [5], RetinaNet [6], Cascaded R-CNN [7], fully connected one-stage (FCOS) detector [8] and FSAF [9], where two sub-nets are used to perform the two tasks separately. In these detectors, however, there is a lack of explicit interactions between the two sub-nets during training.

Recently, the information from the results of localization sub-net is entangled with a new branch incorporating an intersection-over-union (IoU) predictor in [12], and a Kullback-Leibler divergence branch in [13], in order to improve the detection performance, where the ground truth localization results are used as the training labels.

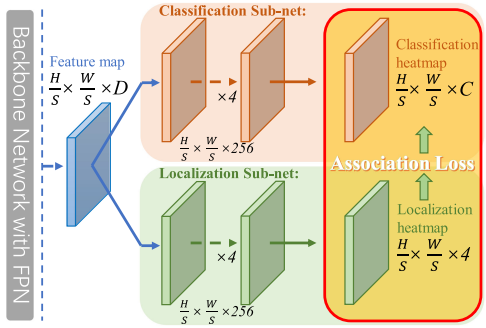


Fig. 1. The architecture of the two sub-nets used in FCOS and the proposed association loss to link the two sub-nets, where the information from the two sub-nets is tied during training to capture their intrinsic relation to improve object detection performance.

Motivated by this idea, here we present a novel association loss for object detection, by including a proxy square error (PSE) loss. Firstly, the localization scores in terms of IoUs are employed by the association loss as additional labels for calculating the classification scores. Then the association loss is integrated with the traditional classification loss in the classification sub-net, which is trained to obtain a joint distribution function between localization and classification scores. Our method is different from a recent study in [10] in the sense that we entangle the information from the two sub-nets with a new loss function, while in [10], a fused feature map is shared between the two sub-nets.

The proposed method offers the following advantages. (1) The classification scores can be associated with the localization scores, which improves the detection accuracy by choosing the best result from several candidate results for the same object according to the classification scores regardless of their localization accuracy, as in [14]. (2) The proposed PSE loss is easy to implement, which can minimize the difference between these two scores without modifying the network structure or training conditions. Our proposed method is evaluated on the MS-COCO dataset, and is shown to improve detection performance over FCOS where the traditional classification loss is used.

II. Preliminary

Our method is based on a recent CNN based detector, FCOS [8], which is a computationally efficient and anchor-free method, with a light-weight structure of two sub-nets, shown in Fig. 1, and discussed next. In the original FCOS, an additional branch has been used to decrease the classification scores predicted from the patches that are not from the center of an object [8]. We omit this branch and use the two sub-nets

Manuscript received April 19, 2020; revised July 10, 2020; accepted July 24, 2020. Date of publication July 31, 2020; date of current version August 28, 2020. This work was supported by the Fundamental Research Funds for the Central Universities under Grant 3072020CFT0602 and in part by the National Natural Science Foundation of China under Grant 61806018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vasileios Mezaris. *(Corresponding author: Jian Guan.)*

Dongli Xu and Jian Guan are with the Group of Intelligent Signal Processing (GISP), College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China (e-mail: dongliixu@gmail.com; j.guan@hrbeu.edu.cn).

Pengming Feng is with the State Key Laboratory of Space-Ground Integrated Information Technology, CAST, Beijing 100095, China (e-mail: p.feng.cn@outlook.com).

Wenwu Wang is with Centre for Vision Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: w.wang@surrey.ac.uk).

Digital Object Identifier 10.1109/LSP.2020.3013160

structure, similar to those in [3]–[7], [9]. Let $\mathbf{G} \in R^{H\times W\times 3}$ be an input image with width W, height H and three color channels. Let $\mathcal{F}_{\text{FPN}}$ be the backbone network with feature pyramid net (FPN) [16] which is used to extract n feature maps as follows

$$\mathbf{F}_0, \mathbf{F}_1, \ldots, \mathbf{F}_j, \ldots, \mathbf{F}_n = \mathcal{F}_{\text{FPN}}(\mathbf{G}) \tag{1}$$

where $\mathbf{F}_j \in R^{\frac{H}{S}\times\frac{W}{S}\times D}$ denotes the feature map at the j-th layer, $j \in \{0, 1, \ldots, n\}$. Here, S is the output stride, and D is the dimensional depth of $\mathbf{F}_j$. Hence, the objective heatmaps of the classification sub-net $\mathcal{F}_{cls}$ and the localization sub-net $\mathcal{F}_{loc}$ can be obtained by projections on each $\mathbf{F}_j$ as follows

$$\widehat{\mathbf{P}}_j = \mathcal{F}_{cls}(\mathbf{F}_j) \tag{2}$$

$$\widehat{\mathbf{B}}_j = \mathcal{F}_{loc}(\mathbf{F}_j) \tag{3}$$

where $\widehat{\mathbf{P}}_j \in [0,1]^{\frac{H}{S}\times\frac{W}{S}\times C}$ and $\widehat{\mathbf{B}}_j \in R^{\frac{H}{S}\times\frac{W}{S}\times 4}$ are the classification and localization heatmap, respectively. Here, $\widehat{\ }$ is used to denote the predictions, C is the number of categories, and there are 4 regression distances (l, t, r, b) from a location (x, y) on $\mathbf{F}_j$ to the four edges of a bounding box centered at the object position.

Following [8], a new position $(\lfloor S\rfloor + xS, \lfloor S\rfloor + yS)$ can be obtained by mapping the location (x, y) onto the input image, where $\lfloor S\rfloor$ takes the largest integer no greater than S. If this position falls into any ground truth bounding box, then (x, y) is considered as a positive sample, and denoted by an indicator $\mathbb{1}_{j,x,y} = 1$, otherwise $\mathbb{1}_{j,x,y} = 0$. Let $\hat{p} \in [0, 1]$ be the prediction from a binary classifier, and $p \in \{0, 1\}$ be the ground truth label for a certain class. Suppose the loss L_{bcls} for a binary classifier is the focal loss L_{focal} [6], expressed as follows

$$L_{bcls} = L_{focal}(\hat{p}, p) = \begin{cases} -\alpha(1-\hat{p})^{\gamma}\log(\hat{p}) & \text{if } p = 1 \\ (\alpha-1)(\hat{p})^{\gamma}\log(1-\hat{p}) & \text{if } p = 0 \end{cases} \tag{4}$$

where α and γ are adjustable parameters, and the use of γ in $(1-\hat{p})^{\gamma}$ and $(\hat{p})^{\gamma}$ can help mitigate the sample imbalance problem [6]. As for localization, let $\hat{I}$ be the IoU for a positive sample, then the IoU loss L_{IoU} [11] is defined as $L_{IoU}(\widehat{I}) = -ln(\widehat{I})$. With N_{pos} positive location samples, the training objectives L_{cls} and L_{loc} for the classification sub-net and the localization sub-net can be expressed respectively as follows

$$L_{cls} = \frac{1}{N_{pos}} \sum_j \sum_{x,y} \sum_c L_{focal}(\widehat{P}_{j,x,y,c}, P_{j,x,y,c}) \tag{5}$$

$$L_{loc} = \frac{\lambda}{N_{pos}} \sum_j \sum_{x,y} \mathbb{1}_{j,x,y} L_{IoU}(\widehat{I}_{j,x,y}) \tag{6}$$

where (x, y) is the location on the heatmap $\widehat{\mathbf{P}}_j$ and $\widehat{\mathbf{B}}_j$, c is the index of the class among the C classes. $\widehat{P}_{j,x,y,c} \in [0, 1]$ and $P_{j,x,y,c} \in \{0, 1\}$ are the prediction and label for classification, respectively. λ is an adjustable weight [8], and $\widehat{I}_{j,x,y}$ is the IoU of the same position calculated as

$$\widehat{I}_{j,x,y} = \frac{T_{j,x,y}}{D_{j,x,y}} \tag{7}$$

where $T_{j,x,y}$ is the area of the overlap region between the predicted bounding box and the ground truth bounding box, $D_{j,x,y}$ is the area of the union region of the predicted bounding box and the ground truth bounding box, where the predicted bounding box can be obtained by using (x, y) and 4 distance $\widehat{\mathbf{B}}_{j,x,y} \in R^{1\times 1\times 4}$. Here, the probability prediction $\widehat{\mathbb{P}}_{j,x,y} = \max_c\{\widehat{P}_{j,x,y,c}\}$ and the IoU $\widehat{I}_{j,x,y}$ are used as the classification score and the localization score, respectively.

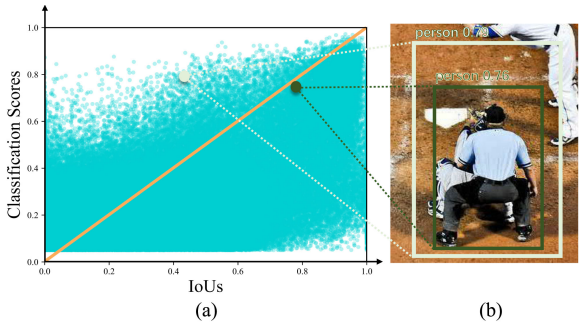


Fig. 2. (a) Scatter plots of the classification score $\widehat{\mathbb{P}}$ versus localization score $\widehat{I}$ for the candidate objects at different pixel positions in the entire test data. Each small dot represents the scores for a candidate position (x, y). The big green dots are used to highlight the inconsistency between $\widehat{\mathbb{P}}$ and $\widehat{I}$, i.e. those dot more distant from the 45 degree anti-vertical yellow line. The scores at those dots may lead the algorithm to the incorrect detection results, e.g. by picking up the predicted bounding box with a higher classification score as the detection result. (b) The light and dark green box have a classification score $\widehat{\mathbb{P}} = 0.79$ and 0.76, and a localization score $\widehat{I} = 0.43$ and 0.77, respectively. The FCOS selected the light box as the detection result due to its slightly higher classification score. However, the dark green box is closer to the ground truth with a higher localization score, despite its relatively lower classification score.

In this baseline method [8], as well as in detectors [3]–[7], [9], the only explicit information shared between the two sub-nets is the input positions on the feature maps. The two sub-nets are trained via two different losses, as a result, the localization scores (i.e. IoUs $\widehat{I}$) obtained from the candidate bounding boxes may not be correlated with the classification score $\widehat{\mathbb{P}}$, as demonstrated in Fig. 2(a), which shows the scatter plots of the classification versus IoU scores for the candidate bounding boxes from the entire test data. It can be seen that the scores for many bounding boxes (such as the light big dots highlighted) are distributed far from the anti-diagonal yellow line. As a result, the algorithm may give incorrect detection results by picking up the candidate bounding box with higher classification scores rather than those with higher localization scores, as exemplified in Fig. 2(b). To address this limitation, we present a new method for associating the localization result with the classification result as detailed next.

III. Association Loss for Object Detection

We propose an association loss to enhance the relation between the localization and classification sub-net, and use this loss to regularize the traditional classification loss to improve the detection performance.

A. PSE Loss

The association loss is defined as a proxy squared error (PSE), as follows

$$L_{\text{PSE}}(\hat{p}, \hat{I}) = \begin{cases} \beta(\hat{p} - k\hat{I})^2 & \text{if } p = 1 \\ 0 & \text{if } p = 0 \end{cases} \tag{8}$$

where β is an adjustable weight, and $\hat{I}$ denotes the localization score, i.e., IoU calculated at the same position in the localization sub-net. Here we use $\hat{p} = k\hat{I}$ with a slope k to denote the linear

relationship between the two scores, where k is adjustable. With this loss, the distance between the two scores can be minimized for a positive sample. For negative samples, the original label (i.e. p) in Eq. (4) is equal to $\hat{I}$ which means the distance is already minimized by the focal loss, hence there is no need to use the PSE loss. As an alternative, the PSE loss can be replaced by a proxy cross-entropy (PCE) loss given as $L_{\text{PCE}}(\hat{p}, \hat{I}) = -\beta log(1 - |\hat{p} - k\hat{I}|)$ if $p = 1$. Incorporating the association loss (e.g. PSE, likewise for PCE), the loss L_{bcls} for a binary classifier can be modified as

$$L_{bcls} = L_{focal} + L_{\text{PSE}} \tag{9}$$

B. An Analysis of the Association Loss

As described in Eq. (4), the loss function used in the traditional classification sub-net for a binary classifier only has one label as well as one regression value. For the convenience of discussion, we set $\gamma = 0$, then L_{focal} is reduced to the well-known cross entropy loss L_{CE} which can provide a similar cumulative absolute loss for positive samples [6]. Let o be the direct output of a classifier before the last sigmoid layer, i.e. $\hat{p} = sigmoid(o)$. For a positive sample, the gradient of L_{CE} with respect to o can be denoted as

$$\frac{\partial L_{\text{CE}}}{\partial o} = \alpha(\hat{p} - 1) \qquad \text{if } p = 1 \tag{10}$$

Using the PSE loss, the gradient of L_{bcls} with respect to o can be calculated as

$$\frac{\partial L_{bcls}}{\partial o} = (\hat{p} - 1)(\alpha + 2\beta\hat{p}^2 - 2\beta k\hat{I}\hat{p}) \quad \text{if } p = 1 \tag{11}$$

Compared with the original gradient in Eq. (10), the factor α is modified to $\alpha + 2\beta\hat{p}^2 - 2\beta k\hat{I}\hat{p}$, which determines the sign of the loss. If $\hat{p}$ is much higher than $\hat{I}$, this factor will be negative and we will have a negative calibration on $\hat{p}$, which will drive this factor towards 0, and the calibration to stop in the minimum of L_{bcls}. This will lead to our objective joint distribution function, denoted as $\hat{p} + \frac{\alpha}{2\beta\hat{p}} = k\hat{I}$. However, there is still a gap $\frac{\alpha}{2\beta\hat{p}}$ between this function and $\hat{p} = k\hat{I}$. Once the classification score $\hat{p}$ goes to 0, the gap will be enlarged. When $k\hat{I}$ is greater than 1, the learning is accelerated due to the relatively higher gradient value, however, the minimum of L_{bcls} will remain the same, since the factor $(\hat{p} - 1)$ will be 0 when $\hat{p}$ is close to 1. In this way, the classification scores and localization scores are better correlated, thus enabling better localization results to be selected in terms of the classification scores. Therefore, the proposed new loss offers an improved joint distribution function, as compared with the traditional classification loss.

C. Implementation and Visualization

As shown in Fig. 1, the association loss is used to regularize the classification loss in the classification sub-net, and thus will not affect the localization sub-net during training. To implement the PSE loss, the training objective L_{cls} for the classification sub-net can be updated as follows

$$L_{cls} = \frac{1}{N_{pos}} \sum_j \sum_{x,y} \sum_c L_{focal}(\widehat{P}_{j,x,y,c}, P_{j,x,y,c}) + \frac{1}{N_{pos}} \sum_j \sum_{x,y} \mathbb{1}_{j,x,y} L_{\text{PSE}}(\widehat{\mathbb{P}}_{j,x,y}, \widehat{I}_{j,x,y}) \tag{12}$$

An alternative approach to associate the classification with the localization score is to use a loss function based on Pearson correlation coefficient (PCC) [15] [17] which, however, may not be accurate for small batches and noisy measurements.

As shown in Fig. 3, we visualize the 3D plot of L_{focal} and L_{PSE} with respect to the classification and localization scores. In Fig. 3(e), there is a valley caused by 0 absolute gradients. The regression stops in the valley that is the same as the minimum of L_{bcls}. This valley can represent objective joint distribution function, since the gradients on two sides of the valley are positive and negative, respectively. However, due to the gap we mentioned above, this valley cannot precisely represent $\hat{p} = \hat{I}$. In this letter, we simply increase β to make the gap close to 0. After increasing the weight for the PSE loss, the valley in Fig. 3(f) approaches $\hat{p} = \hat{I}$. Compared with Fig. 3(d), we can see that the original regression objective $\hat{p} = 1$ is changed to $\hat{p} = \hat{I}$, approximately. However, the gradients become arbitrarily small as $\hat{p}$ approaches 1, which could be modified to further improve the classification loss, but is out of the scope of this work.

The proposed PSE loss is designed for the classification sub-net, which is, however, not suitable for the localization sub-net. This is because the localization task only uses positive samples for regression. With the PSE loss, some positive samples will become negative if the IoU is greater than the classification score, thus decreasing the localization performance. It would be interesting to design an association loss for the localization sub-net.

IV. Experiments and Results

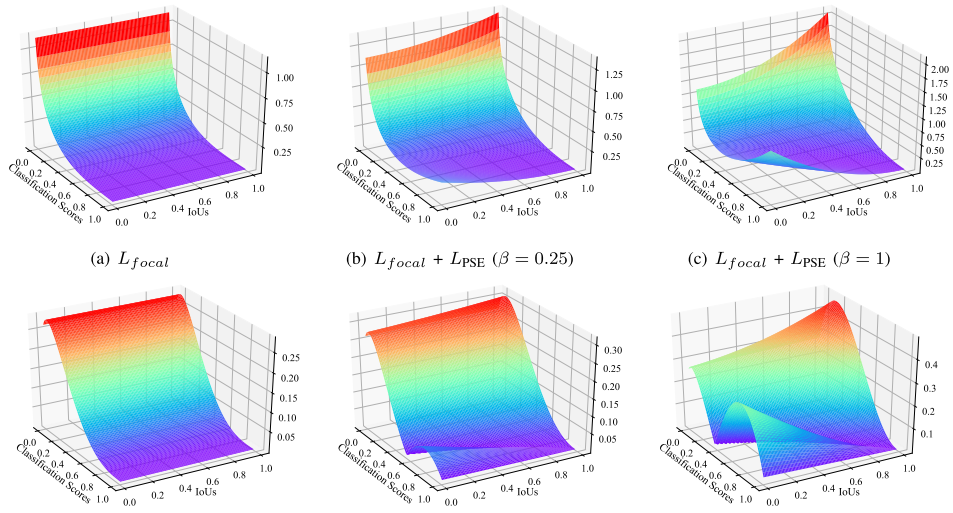
Experiments are conducted on the MS-COCO dataset [18] to demonstrate the performance of our association loss, where average precision (AP) and average recall (AR) are employed as performance metrics [18], e.g., AP_{50} denotes AP at true positive predictions with threshold IoU $= 0.5$, AP_{S}, AP_{M} and AP_{L} denote AP for small, middle and large objects, respectively, and AR_1 denotes AR given 1 detection per image. The data `trainval35k(115k images)` is used for training and `minival(5k images) is used as validation.` The backbone network is Res-Net-50 [19], the implementation is based on Pytorch 1.0, CUDA 10, maskrcnn-benchmark[1] and FCOS. All hyper-parameters are set to be default as in FCOS unless specified:[2] the initial learning rate in SGD is set as 0.01 and divided by 10 at 60k and again at 80k iterations. The weight decay is 0.0001 and the momentum is 0.9. The batchsize is set as 16 and the warms-up scheme is applied for the first 500 iterations. We set $\alpha = 0.25$, $\gamma = 2$ in the focal loss and $\lambda = 1$ following [6], [8].

A. Ablation Study

We evaluate our method on the FCOS (without the centerness branch mentioned earlier). Here we set $\beta = 1, k = 1$. The result is given in Table I. Here, the baseline classification sub-net loss L_{cls} is a traditional focal loss. By adding our association loss (PSE) to L_{cls}, we can improve the AP and AP_{75} by 1.6 and 2.4, respectively. This demonstrates that our proposed PSE loss improves detection accuracy. However, the PSE loss does not increase the AP_{S} for small objects as much as AP_{L} for large

[1][Online]. Available: https://github.com/facebookresearch/maskrcnn-benchmark

[2][Online]. Available: https://github.com/tianzhi0549/FCOS



(a) L_{focal} (b) L_{focal} + L_{PSE} ($\beta = 0.25$) (c) L_{focal} + L_{PSE} ($\beta = 1$)

(d) Absolute Values of Gradients of (a) (e) Absolute Values of Gradients of (b) (f) Absolute Values of Gradients of (c)

Fig. 3. The 3D version of losses for positive samples: (a) L_{focal}, (b) L_{focal} + L_{PSE} ($\beta = 0.25$), (c) L_{focal} + L_{PSE} ($\beta = 1$). (d) (e) (f) are absolute values of gradients of (a) (b) (c), respectively. An important theme to notice: these absolute gradients on the two sides of a valley in (e) (f) have opposite signs with each other. Here we set $\alpha = 0.25$, $\gamma = 2$ and $k = 1$. We plot these losses and gradients from Classification Scores $= 0.001$ to 1 and IoUs of same range.

TABLE I
THE CONTRIBUTION OF ASSOCIATION LOSS TO DETECTION RESULTS ON FCOS (WITHOUT CENTERNESS). HERE WE SET $\beta = 1$, $k = 1$

L_{cls}	AP	AP_{50}	AP_{75}	AP_{90}	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}
Focal Loss [6]	33.8	52.8	35.4	12.1	19.6	38.7	44.2	30.2	47.9	52.1
Focal Loss + Association Loss (PCE)	35.0	53.1	37.4	13.4	19.4	39.4	**47.3**	30.3	47.9	51.5
Focal Loss + Association Loss (PSE)	$\mathbf{35.4}^{+1.6}$	$\mathbf{53.7}^{+0.9}$	$\mathbf{37.8}^{+2.4}$	$\mathbf{13.6}^{+1.5}$	$\mathbf{20.2}^{+0.6}$	$\mathbf{39.8}^{+1.1}$	$47.1^{+2.9}$	$\mathbf{30.9}^{+0.7}$	$\mathbf{48.8}^{+0.9}$	$\mathbf{52.7}^{+0.6}$

TABLE II
ASSOCIATION LOSS ANALYSIS BETWEEN TWO SCORES THAT BELONG TO ALL CANDIDATE BOXES (NEARLY 250K) OF 80 CLASSES AND ALL IMAGES IN `Minival`

Methods	PCC	SCC	KCC
w/o Association Loss	0.528	0.433	0.294
w/ Association Loss (PSE)	**0.534**	**0.438**	**0.298**

TABLE III
PERFORMANCE EVALUATION OF THE PSE LOSS FOR DIFFERENT VALUES OF k (FIXING $\beta = 1$) AND β (FIXING $k = 1$) RESPECTIVELY

k	AP	AP_{50}	AP_{75}	β	AP	AP_{50}	AP_{75}
0.8	35.1	53.9	37.2	0.25	34.4	53.1	36.6
1	**35.4**	**53.7**	**37.8**	0.8	35.1	53.5	37.4
1.25	34.8	53.2	37.1	1	**35.4**	**53.7**	**37.8**
1.5	34.4	52.9	36.5	1.25	35.3	53.5	37.6

objects. This is probably because there are a large number of candidate boxes for a single large object, while there are much less candidate boxes for a single small object. We have also tested the PCE loss and the results show that the PCE loss improves AP by 1.2 but the improvement is not as much as that by the PSE loss.

To better understand the effect of incorporating the association loss, we analyze the distribution of two scores of the final detection results with three correlation coefficients: PCC, Spearman rank correlation coefficient (SCC) and Kendall rank correlation coefficient (KCC). Here we set $\beta = 1$, $k = 1$. The analysis is given in Table II, here the w/o Association Loss denotes our detector is trained without any association loss, w/ Association Loss (PSE) denotes our detector is trained with the PSE loss. The results show that with our method the performances of the two sub-nets can be better correlated.

In order to find the relationship between $\hat{p}$ and $\hat{I}$, experiments with different slope k (fixing $\beta = 1$) are performed. The results are given in Table III. It can be seen that $k = 1$ gives better results than other choices. It is worth noting, however, that k needs to be tuned empirically for the dataset at hand. We also conduct experiments with different weight β (fixing $k = 1$) to evaluate its influence on the detection performance. The results are also given in Table III. We can observe that better result can be obtained when β is greater than the weight α in the focal loss (i.e. $\beta = 1, \alpha = 0.25$). This is because we can get a more accurate objective joint distribution function by increasing the weight β of the PSE loss as we discussed in Section III-C. However, when β is greater than 1, the performance will no longer be improved (e.g. $\beta = 1.25, \alpha = 0.25$), as this may re-introduce the sample imbalance problem addressed by the focal loss [6].

V. CONCLUSION

We presented a novel association loss to entangle the localization scores with the classification scores, and applied it to the classification sub-net in a two sub-nets based CNN detector. Our proposed loss can transfer information between two sub-nets and correlate the classification scores with localization scores. Experiments conducted demonstrate that our method can better correlate the two sub-nets, and hence improve the detection accuracy over the usage of traditional classification loss.

REFERENCES

[1] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[5] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[6] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[7] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
[8] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
[9] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 840–849.
[10] J. U. Kim, S. T. Kim, E. S. Kim, S. K. Moon, and Y. M. Ro, "Towards high-performance object detection: Task-specific design considering classification and localization separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process*, 2020, pp. 4317–4321.
[11] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 516–520.
[12] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 784–799.
[13] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2888–2897.
[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
[15] J. Benesty, J. Chen, and Y. Huang, "On the importance of the Pearson correlation coefficient in noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 757–765, May 2008.
[16] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
[17] J. Benesty, J. Chen, and Y. Huang, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, Berlin, Germany: Springer, 2009.
[18] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.