Semi-DETR: Semi-Supervised Object Detection with Detection Transformers (*Supplementary Document*)

Jiacheng Zhang[1,2*] Xiangru Lin[2*] Wei Zhang[2] Kuo Wang[1] Xiao Tan[2]
Junyu Han[2] Errui Ding[2] Jingdong Wang[2] Guanbin Li[1,3†]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Department of Computer Vision Technology (VIS), Baidu Inc., China
[3]Research Institute, Sun Yat-sen University, Shenzhen, China
{zhangjch58, wangk229}@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn
{linxiangru,zhangwei99,tanxiao01,hanjunyu,dingerrui,wangjingdong}@baidu.com

1. Extended Details of Related Work

Semi-Supervised Object Detection. In SSOD, Pseudo Labeling [17] and Consistency-based Regularization [6], [12] are two commonly used methods. As an early SSOD work, STAC [14] proposed a basic multi-stage training framework to combine pseudo labeling and consistency training. To simplify the multi-stage training process, the end-to-end Teacher-Student framework [15], [10] is proposed, in which the teacher model is updated by exponential moving average (EMA) from the student model, and generates pseudo labels on the unlabeled images in an online manner. Under this framework, a significant amount of research is proposed to improve the quality of pseudo labels [10, 18, 20, 25]. Among them, Unbiased Teacher [10] replaces the Cross-Entropy loss with Focal Loss to eliminate the class imbalance caused by confirmation bias [1] of pseudo labels. For Consistency-based regularization methods [5, 9], PseCo [9] introduces the feature-level scale consistency by aligning shifted pyramid features of different scale inputs of the same image. Most of these works are based on the two-stage detectors, e.g. Faster RCNN [13], which involves the anchor generator, a complex hand-crafted component. On the other hand, some SSOD methods focus on the one-stage detectors [3, 11, 24]. Among them, DSL [3] proposed the first dense learning-based anchor-free SSOD method with adaptive filtering strategy and uncertainty regularization and achieved state-of-the-art performance. Dense Teacher [24] is proposed to use the dense output predictions from the teacher branch as pseudo labels directly to avoid the annoying threshold selection. Our Semi-DETR is significantly different from previous works: (1) we explored the challenges of the DETR-based object detectors on SSOD, which, to our best knowledge, is the first systematic research endeavor in SSOD; (2) our Semi-DETR method is tailored for the DETR-based detectors, which eliminates the training efficiency caused by bipartite matching with the noisy pseudo labels and presents a new consistency scheme for set-based detectors.

2. Extended Details of Stage-wise Hybrid Matching

Design Details. In Stage-wise Hybrid Matching, we propose to divide the training process into two stages: the one-to-many assignment training in the first stage and the one-to-one assignment training in the second stage. Following the main paper, let us denote the classification score as s and an IoU between the predicted bounding box and the ground truth bounding box as u. We take a high-order combination of the classification score and the IoU as the matching score and its negative version as the matching cost in the first stage:

$$m = s^{\alpha} \cdot u^{\beta} \tag{1}$$

$$\mathcal{C}_{\text{match}}(\hat{y}_i^t, \hat{y}_j^s) = -m_{ij} = s_{ij}^{\alpha} \cdot u_{ij}^{\beta} \tag{2}$$

where $\hat{y}^t$ and $\hat{y}^s$ are the pseudo labels generated by the teacher and the prediction of the student, respectively. The s_{ij} is the classification score of j-th bounding box prediction to i-th ground truth label, and u_{ij} is the IoU between the j-th predicted bounding box and the i-th ground truth box. The higher the matching score, the better the matching quality between the predicted bounding box and the ground truth box, and the lower the matching cost between them. Then, we assigned multiple positive proposals to each

[*] Equally-contributed authors. Work done during an internship at Baidu.

[†] Corresponding author.

pseudo box according to the matching score as follows:

$$\hat{\sigma}_{o2m} = \left\{ \underset{\boldsymbol{\sigma_i} \in C_N^M}{\arg\min} \sum_{j=1}^{M} \mathcal{C}_{\text{match}}\left(\hat{y}_i^t, \hat{y}^s_{\boldsymbol{\sigma_i(j)}}\right) \right\}_{i=1}^{|\hat{y}^t|} . \tag{3}$$

where $|\hat{y}^t|$ is the number of the pseudo labels. C_N^M is the combination of M and N, which denotes that a subset of M proposals is assigned to the pseudo box $\hat{y}_i^t$, and the $\sigma_i(j)$ is the corresponding positive proposal indices. With this assignment strategy, the number of assigned positive proposals for each pseudo label significantly increases, boosting the probability of containing the proposals with higher quality as positive samples and leading to more efficient training. In the implementation, we simply choose the Top-K (K=M) proposals with the largest matching scores for each pseudo box as the positive proposals.

After the first stage of training, the model is capable to produce high-quality pseudo labels with NMS as the post-process. To enjoy the merit of NMS-free detection without sacrificing the detection performance, we propose to conduct one-to-one assignment training with both the labeled data and unlabeled data in the second stage, where the NMS post-process is applied to the unlabeled data to provide high-quality pseudo boxes. The one-to-one assignment, along with the high-quality pseudo boxes, helps the model to gradually reduce the duplicated predictions and finally evolve into an NMS-free end-to-end detector with better performance.

Statistical Analysis. To validate the effectiveness of our method, we first get the positive candidate proposals obtained by one-to-one assignment using pseudo bounding boxes and the positive candidate proposals obtained by one-to-one assignment using corresponding ground-truth bounding boxes, respectively:

$$\begin{aligned} b_i^{o2o} &= A_{o2o}(b_i^{pd}) \\ \hat{b}_i^{o2o} &= A_{o2o}(b_i^{gt}) \end{aligned} \tag{4}$$

where b_i^{pd} and b_i^{gt} is the i-th pseudo box and its corresponding ground-truth box. A_{o2o} means the one-to-one assignment, i.e. bipartite matching, and the b_i^{o2o} and $\hat{b}_i^{o2o}$ are the corresponding assigned positive proposals. We then calculate the IoU between these two assigned positive proposals:

$$I_i^1 = IoU(b_i^{o2o}, \hat{b}_i^{o2o}) \tag{5}$$

The IoU value I_i^1 represents the quality of the assigned candidate proposal. The larger the IoU, the closer the assigned positive candidates are to the target object and the better the quality. As a comparison, we get the assigned positive proposals by one-to-many assignment for the i-th pseudo box:

$$\boldsymbol{b_i^{o2m}} = \{b_{i1}^{o2m}, b_{i2}^{o2m}, ..., b_{im}^{o2m}\} = A_{o2m}(b_i^{pd}) \tag{6}$$

where A_{o2m} is our one-to-many assignment strategy, $\boldsymbol{b_i^{o2m}}$ is the multiple assigned positive proposals for i-th pseudo box. To verify whether there are positive proposals with higher quality contained in the proposal set obtained by the one-to-many assignment strategy, we calculate the max IoU of these multiple positive proposals and the positive proposal $\hat{b}_i^{o2o}$:

$$I_i^2 = Max(\{IoU(b_{i1}^{o2m}, \hat{b}_i^{o2o}), ..., IoU(b_{im}^{o2m}, \hat{b}_i^{o2o})\}) \tag{7}$$

Then, we compare the IoUs I_i^1 and I_i^2, and the results are shown in Fig. 1. It can be found that the multiple positive proposals obtained by our one-to-many assignment strategy clearly contain proposals with higher quality than the proposal obtained by the one-to-one assignment strategy. This result demonstrates that a number of proposals with poor quality are assigned as positive samples due to inaccurate pseudo boxes in the one-to-one assignment, while the correct positive candidate proposals with higher quality are forcibly assigned as negative samples, which finally causes inefficient training. Our proposed Stage-wise Hybrid Matching applies the one-to-many assignment strategy in the first stage, which enables the proposals with higher quality mistakenly assigned as background to have the opportunity to be optimized. With the modified loss function, the potential positive proposals can be utilized to guide the model convergence while the impact of the proposals with low quality can also be eliminated at the same time.

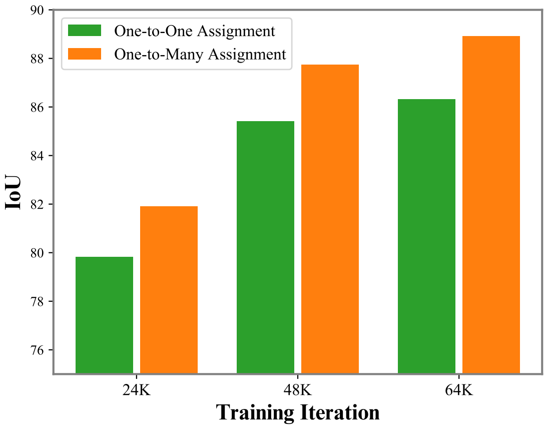


Figure 1. The investigation of the quality of the assigned positive proposals obtained by the one-to-one assignment and one-to-many assignment.

Effectiveness Analysis. As presented in Fig. 4, compared with the one-to-one assignment with Bipartite Matching [2], the proposed Stage-wise Hybrid Matching greatly improves the training efficiency of the first stage thanks to the multiple assigned positive proposals. More importantly, the performance improvement becomes more prominent when the number of labeled data gets more scarce, e.g., 1%, which demonstrates the superiority of our method.

Furthermore, we compare different alternative one-to-many assignment strategies in our method. Specifically, we replace the one-to-many assignment strategy used in our method with the Max-IoU [13], ATSS [23], and SimOTA [4], respectively. We conduct these experiments with Semi-DETR with DINO [22], and all the models are trained for 60K iterations. The results are shown in Tab. 1. Interestingly, although designed to assign multiple positive proposals, both Max-IoU and ATSS do not perform well in DETR-based detectors. We visualize the assignment results of these assignment strategies in Fig. 3. When applying the Max-IoU assignment strategy, we observe that only a few ground truth boxes own lots of duplicated positive proposals, and most ground truth boxes have no positive assigned proposals. We suspect the main reason is that the learnable object queries are unable to always guarantee enough IoU with each ground truth box, and constantly changed during the training. Unlike the fixed anchor box prior, the predicted proposals of the object queries easily cluster around a particular ground truth box, and finally leads to duplicated positive proposals, which is not helpful for the training. However, further discussion is beyond this paper. As a comparison, the ATSS assignment strategy generates a few positive proposals for each ground truth box. However, the number of positive proposals obtained by the ATSS for each ground truth box is still limited. This is because the ATSS only considers the IoU during the assignment, and the adaptive IoU thresholds obtained by the ATSS are so high that most of the possible high-quality proposals are filtered out. Different from the Max-IoU and ATSS, both SimOTA [4] and our proposed method achieve much better performance, which benefits from (1) the ranking-based one-to-many assignment strategy via the top-K operation to ensure enough positive proposals for each ground truth box, and (2) the ranking criteria considering both the classification score and IoU score, which can generate various positive proposals as shown in 3.

Table 1. Performance comparisons among different one-to-many assignment strategies. Baseline means the one-to-one assignment with bipartite matching. All the models are trained for 60K iterations.

Method	mAP	AP_{50}	AP_{75}
Baseline	40.2	56.5	43.4
Max-Iou	11.4	15.0	12.1
ATSS	18.7	30.5	18.9
SimOTA	42.5	**59.9**	45.2
Ours	**42.8**	59.8	**46.0**

3. Extended Details of Cross-view Query Consistency

We conduct experiments to validate the effectiveness of our cross-view query consistency. We aim to answer the following two questions:

1. Whether the cross-view queries necessary? Can we replace the cross-view queries with single-view queries?
2. Whether the RoI features in the cross-view queries really matter? What about conducting consistency training without incorporating these features into the consistency queries?

To answer these two questions, we conduct two experiments as follows:

Exp-1: We construct the consistency queries by the RoI features within each view separately. And then, we perform the query decoding in the teacher and student models individually. Finally, we impose the consistency constraint on the consistency queries decoding embedding of the teacher and student models. The overview is presented in Fig. 5(b). As shown in Tab. 2, compared with our proposed cross-view query consistency, when replacing the cross-view queries with single-view queries, the performance shows a 0.5 mAP drop. It confirms the importance of the cross-view queries in our consistency scheme. The possible reason for the effectiveness of these cross-view queries is that these queries provide information about the object from another view, which encourages learning the semantic invariance during decoding and leads to better performance.

Exp-2: As shown in Fig. 5(c), we construct the consistency queries directly based on the positional embedding of the pseudo boxes without the RoI features of corresponding pseudo boxes. The difference between this scheme and DN-DETR [8] is that we do not add the noise into the pseudo boxes before obtaining the positional embedding. The results are presented in Tab. 2. After removing the RoI features during the construction of the consistency queries, the performance greatly decreased to 42.7(-0.8) mAP, which demonstrates the necessity of the RoI features in consistency queries. The reason for this performance degeneration is that the positional embedding of the pseudo boxes does not have strong priors to guarantee the correspondence between the consistency queries input and their corresponding output prediction, which increases the learning difficulty of the consistency training. As a comparison, we take the RoI features from different views as the strong semantic guidance during the decoding and ensure the final decoder embedding is relevant to the input consistency queries, which eventually leads to the success of the consistency training.

Table 2. Performance comparisons of different variants of Cross-view Query Consistency

Method	mAP	AP_{50}	AP_{75}
Exp-1	43.0	59.3	46.3
Exp-2	42.7	58.9	46.0
Ours	**43.5**	**59.7**	**46.8**

4. Extended Details of the Cost-based Pseudo Label Mining

Design Details. Concretely, we take two steps to generate the pseudo boxes for consistency training with a good trade-off between precision and recall. First, for each unlabeled image, we calculate the mean μ and variance σ of the confidence scores of the detection results. Then, we take the threshold $\tau_1 = \mu + \sigma$ to filter and get the initial pseudo boxes. For the second step, we perform the bipartite matching with these initial pseudo boxes and the student model predicted proposal boxes, and record the matching cost of each pseudo box. We collect the matching cost of the initial pseudo boxes within a batch and show the distribution of the matching cost in Fig. 2. Obviously, the distribution of the matching costs presents a bimodal distribution. To this end, we propose to model the cost distribution with a Gaussian Mixture Model(GMM) consisting of two Gaussian distributions as follows:

$$P(c|\theta) = w_r \mathcal{N}_r(c, \mu_r, \sigma_r) + w_u \mathcal{N}_u(c, \mu_u, \sigma_u) \quad (8)$$

where the $P(c|\theta)$ means the probability of matching cost value c, θ is the parameters of the GMM model. $\mathcal{N}_r(c, \mu_r, p_r)$ represents the cost distribution of reliable pseudo boxes with lower matching cost and $\mathcal{N}_u(c, \mu_u, p_u)$ represents the cost distribution of unreliable pseudo boxes with higher matching cost. w_r and w_u represent the blending weights of these two distributions, μ_r(or μ_u), and σ_r(or σ_u) represent the means and variances of these two distributions, respectively. The fitting process can be solved by the Expectation-Maximization (EM) algorithm [19]. Then, we set the threshold τ_c as the cost with the highest probability of being the reliable pseudo boxes.

$$\tau_c = \arg\max_c P_{reliable}(c|c, \theta) \quad (9)$$

The bounding boxes with matching costs less than τ_c are regarded as reliable pseudo boxes and are retained for cross-view query consistency learning. As shown in Fig. 6, this pseudo label mining method successfully mines more pseudo bounding boxes which is reliable for consistency training from the initial pseudo boxes.

Effectiveness Analysis. In our main paper, we take the fixed classification confidence score $\tau_s = 0.4$ to filter and obtain the pseudo labels for the training of classification and regression. The pseudo labels obtained by cost-based pseudo label mining (CPM) are used for consistency training only. Here, we conduct experiments to test the extension of the CPM to replace this fixed threshold filtering scheme. According to Tab. 3, interestingly, when the pseudo labels from the CPM are utilized to train the classification and regression losses, the detector suffers a clear performance drop (-1.1%). This indicates that these pseudo labels are not suitable for the training of classification and regression. The possible reasons for this performance drop are two-fold: (1) our proposed Cross-view Query Consistency aims to learn semantic feature invariance between different views from unlabeled images, which essentially does not have a strict requirement for the localization accuracy(i.e. high precision) of the pseudo bounding boxes. Meanwhile, the CPM generates more pseudo bounding boxes(i.e. high recall) than that of the fixed threshold filtering scheme, which essentially facilitates the learning of cross-view query consistency. (2) As discussed in the Stage-wise Hybrid Matching part in the main paper, the one-to-one assignment strategy used in DETR-based detectors requires more accurate(i.e. high precision) pseudo labels to effectively supervise the classification and regression learning, otherwise would lead to inefficient training.

5. Data Augmentations

Generally, we follow the data augmentation scheme in Soft-Teacher [18]. We summarize the data augmentations used in our method in Tab. 5. Note that we do not use more advanced data augmentations such as Large Scale Jittering in [18, 24], MixUp [21], and Mosaic in [25], Patch Shuffle in [3]. We believe these data augmentations can further improve our performance, which we leave for future work.

Table 3. Experiments about usage extension of the pseudo labels from Cost-baed Pseudo Label mining(CPM). Cls means classification training and Reg means regression training. Consistency represents the cross-view query consistency.

Method	Cls + Reg	Consistency	mAP
CPM(Ours)		✓	**43.5**
CPM(Extension)	✓	✓	42.4

6. Extended Details of Experiments

Here, we provide more details about the experiments with Deformable DETR [26], i.e. Semi-DETR(Def-DETR). (1) For the COCO Partial benchmark, we train Semi-DETR(Def-DETR) for 180k iterations and the training time of first stage with one-to-many assignment T_1 is set to 120k iterations. Other settings are kept the same with Semi-DETR(DINO). (2) For the COCO-Full benchmark, the total training time is set to 240k iterations, and T_1

is set to 180k iterations. Other settings are kept the same with Semi-DETR(DINO). (3) For the Pascal VOC benchmark, we train Semi-DETR(Def-DETR) for 120k iterations with the training time of first stage T_1 set to 80k iterations. Other settings are kept the same with COCO-Partial benchmark. For all experiments, the confidence threshold is set to 0.4. We utilize Adam [7] with a learning rate of 2e-4 and weight decay of 0.0001, and no learning rate decay scheme is used. The teacher model is updated from the student model through EMA with a momentum of 0.999.

Comparisons to Omni-DETR. Omni-DETR [16] is a DETR-based object detector designed for omni-supervised object detection. It is not designed specifically for SSOD as admitted in their paper, but it is extended to the SSOD task by introducing a simple pseudo-label filtering scheme. Our Semi-DETR is significantly different from Omni-DETR in the following aspects:

(1) **Different motivations for model design**. To perform SSOD, Omni-DETR adopted simple hard thresholding on the confidence scores of the predictions to assign supervised pseudo-labels to unlabeled data, which can be viewed as a simple integration of DETR-based detectors to the general SSOD framework. We conducted an in-depth analysis of this pipeline and identified that the one-to-one assignment strategy leads to training inefficiency due to inaccurate pseudo labels, especially during the early training phase. Besides, the lack of deterministic correspondence between the input query and its prediction output in DETR-based detection framework also hinders the integration of consistency-based regularization which is known to be effective in existing SSOD methods. Consistency-based regularization is therefore not explored in Omni-DETR. Our proposed Semi-DETR alleviates the training inefficiency by combining the one-to-many and one-to-one assignment strategies to provide pseudo-labels of higher quality. Moreover, it introduces a consistency-based regularization scheme powered by a cost-based pseudo label mining method, which enables consistency regularization for DETR-based detectors. In general, compared to Omni-DETR, Semi-DETR is a tailored design for SSOD, and it is an important step forward to extend the study of DETR-based detectors to SSOD.

(2) **Different training strategy**. Omni-DETR follows the complex multi-stage training pipeline of Unbiased-Teacher [10], which requires an extra burn-in stage to pre-train on labeled data and thus **is not essentially end-to-end**. However, Semi-DETR shares the same design philosophy as Soft-Teacher [18] without the need to pre-train with labeled data in advance. Both detectors embrace the benefits of NMS-free post-process, but our proposed Semi-DETR achieves end-to-end detection in both the training and inference phases. **This strengthens our claim that Semi-DETR is the first transformer-based end-to-end semi-supervised object detector**.

(3) **Significant performance improvement**. As discussed in [16], Omni-DETR utilizes Deformable DETR as the base detector for faster convergence. We compare our Semi-DETR with Omini-DETR using the same baseline detectors under different COCO-Partial settings as in Tab. 4. Clearly, Semi-DETR achieves SOTA performance with different detectors, and it is superior to Omni-DETR across all base detectors under different experimental settings.

We present fair comparisons between Omni-DETR and Semi-DETR using different base detectors (i.e., Deformable DETR and DINO) in Tab. 4. **First we must clarify that Omni-DETR actually adopts Deformable DETR as the base detector due to the slow convergence of original DETR**. The performance of Omni-DETR with Deformable DETR is thus directly copied from [16], and we additionally evaluate its performance with DINO. As shown in Tab. 4, Semi-DETR consistently achieves better performance than Omni-DETR across all settings. Moreover, **even armed with DINO as the base detector, our Semi-DETR still outperforms Omni-DETR by clear margins**, which manifests the superiority of our Semi-DETR in terms of performance compared to Omni-DETR.

Table 4. Performance comparisons between Omni-DETR and Semi-DETR with different detectors under COCO-Partial settings.

Method	1%	5%	10%
Omni-DETR(Def-DETR)	18.60	30.20	34.10
Semi-DETR(Def-DETR)	**25.20**	**34.50**	**38.10**
Improvement	+6.60	+4.30	+4.00
Omni-DETR(DINO)	27.60	37.70	41.30
Semi-DETR(DINO)	**30.50**	**40.10**	**43.50**
Improvement	+2.90	+2.40	+2.20

7. More Visualization

Stage-wise Hybrid Matching improves the training efficiency when the pseudo labels are inaccurate by the one-to-many assignment, and makes it able to generate the pseudo labels with higher quality in the second stage. To validate this, we visualize the pseudo labels training with and without the State-wise Hybrid Matching. The results are shown in Fig. 7. It clearly shows that our Stage-wise Hybrid Matching generates better pseudo boxes.

8. Discussion of Limitations

Achieving an end-to-end detection framework without NMS post-processing under DETR-based semi-supervised object detection, while maintaining the performance of a fully one-to-many assignment strategy is a research direction worth exploring. Semi-DETR has demonstrated the effectiveness of combining the one-to-many assignment and the one-to-one assignment strategies at the cost of a performance drop compared to the fully one-to-many assignment strategy. Nevertheless, how to design a better DETR-based SSOD framework that could minimize this performance gap remains an open problem in the research community. We leave it to future work.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[3] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4815–4824, 2022. 1, 4

[4] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3

[5] Qiushan Guo, Yao Mu, Jianyu Chen, Tianqi Wang, Yizhou Yu, and Ping Luo. Scale-equivalent distillation for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14522–14531, 2022. 1

[6] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32, 2019. 1

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[8] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 3

[9] Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Pseco: Pseudo labeling and consistency training for semi-supervised object detection. *arXiv preprint arXiv:2203.16317*, 2022. 1

[10] Yen-Cheng Liu and et al. Unbiased teacher for semi-supervised object detection. In *ICLR 2021*. 1, 5

[11] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9819–9828, 2022. 1

[12] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 1

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 3

[14] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1

[15] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1

[16] Pei Wang and et al. Omni-detr: Omni-supervised object detection with transformers. In *CVPR 2022*. 5

[17] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[18] Mengde Xu and et al. End-to-end semi-supervised object detection with soft teacher. In *ICCV 2021*. 1, 4, 5

[19] Guorong Xuan, Wei Zhang, and Peiqi Chai. Em algorithms of gaussian mixture model and hidden markov model. In *Proceedings 2001 international conference on image processing (Cat. No. 01CH37205)*, volume 1, pages 145–148. IEEE, 2001. 4

[20] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5941–5950, 2021. 1

[21] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2017. 4

[22] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3

[23] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 3

[24] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun. Dense teacher: Dense pseudo-labels for semi-supervised object detection. *arXiv preprint arXiv:2207.02541*, 2022. 1, 4

[25] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4081–4090, 2021. 1, 4

[26] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable trans-

formers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4

Table 5. Data augmentations used in our method. p represents the probability of choosing a certain type of augmentation.

Augmentation	Labeled image training	Unlabeled image training	Pseudo-label generation
Scale Jitter	shortest edge $\in [480, 800]$	shortest edge $\in [480, 800]$	shortest edge $\in [480, 800]$
Solarize Jitter	p=0.25,ratio∈(0,1)	p=0.25,ratio∈(0,1)	-
Brightness	p=0.25,ratio∈(0,1)	p=0.25,ratio∈(0,1)	-
Constrast Jitter	p=0.25, ratio ∈(0,1)	p=0.25, ratio ∈(0,1)	-
Sharpness Jitter	p=0.25, ratio ∈(0,1)	p=0.25, ratio ∈(0,1)	-
Translation	-	p=0.3, translation ratio∈(0,1)	-
Rotate	-	p=0.3,angle∈(0,30°)	-
Shift	-	p=0.3,angle∈(0,30°)	-
Cutout	num∈(1,5),ratio∈(0.05,0.2)	num∈(1,5),ratio∈(0.05,0.2)	-

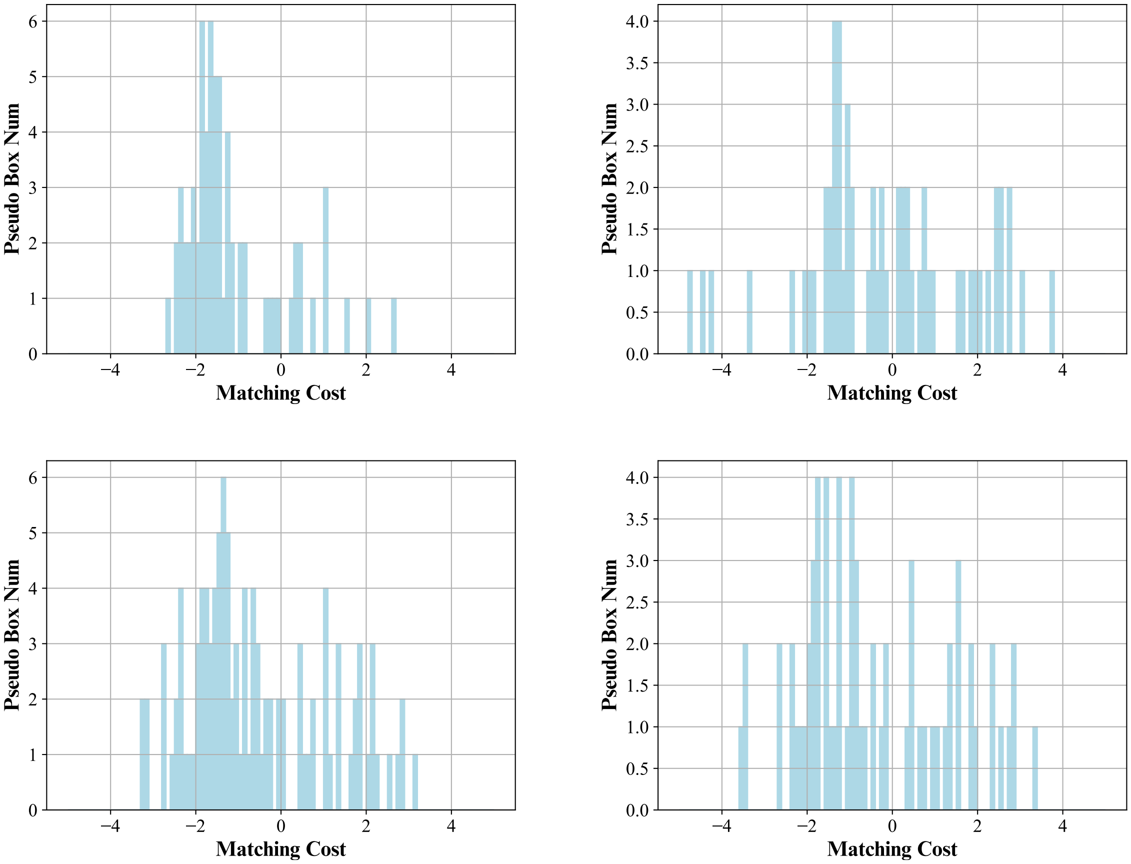
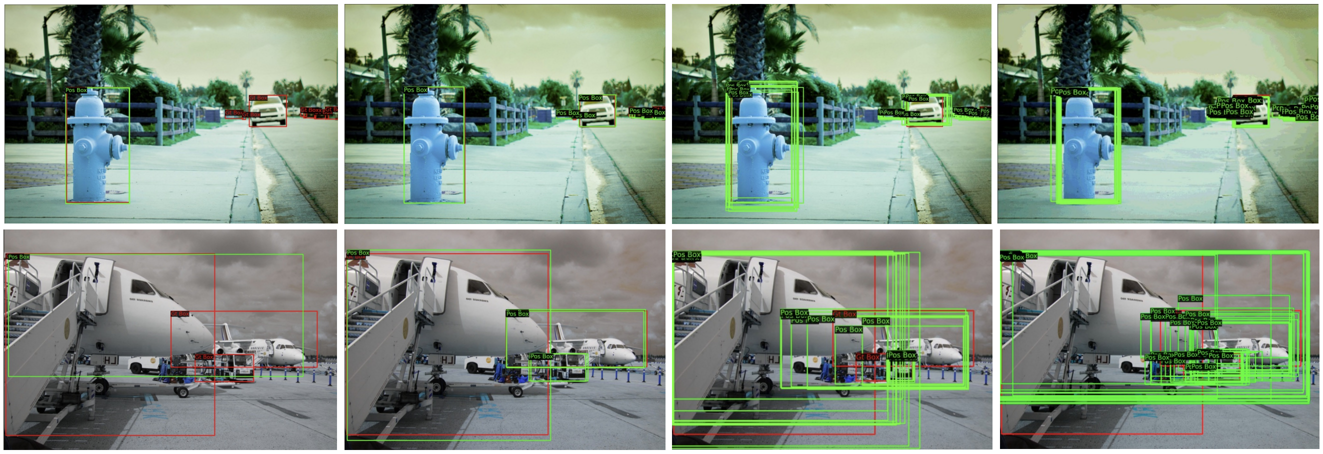
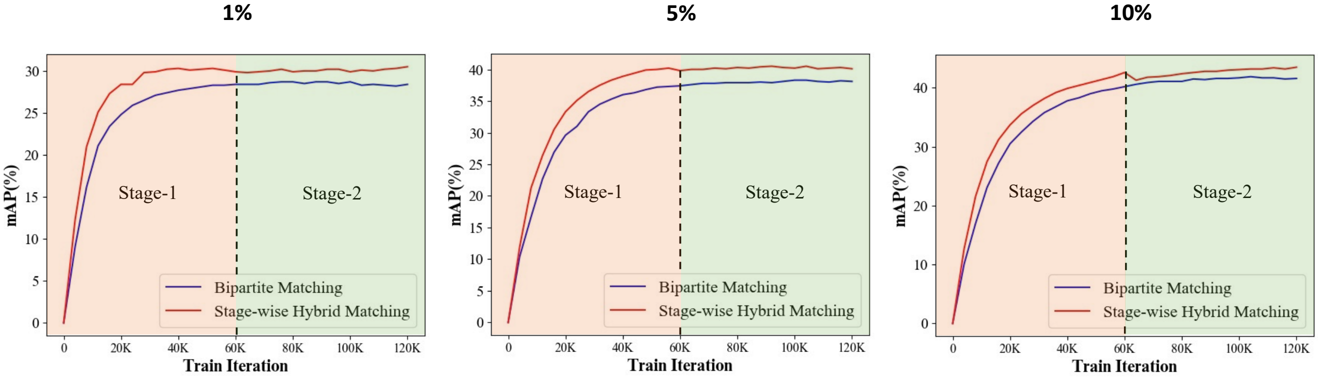


Figure 2. The distribution of the matching cost of the initial pseudo boxes within a random batch.



(a) (b) (c) (d)

Figure 3. Qualitative results of assigned positive proposals of different one-to-many assignment strategies. (a) Max-IoU(IoU Threshold 0.5) (b) ATSS (c) SimOTA (d) Ours. Compared to Max-IoU, ATSS, and SimOTA, our method obtains more positive proposals for each ground truth bounding box. Note that the ground truth bounding boxes are in red, and the assigned positive predicted bounding boxes are in green.



(a) (b) (c)

Figure 4. The training efficiency comparisons between the proposed **State-wise Hybrid Matching**(one-to-many assignment in stage-1 and one-to-one assignment in stage-2) and the original **Bipartite Matching**(one-to-one assignment) under different labeled data ratios on the COCO dataset. The area in orange and green represent the first stage and second stage in the Stage-wise Hybrid Matching, respectively. Our Stage-wise Hybrid Matching greatly improves the training efficiency, especially when labeled data are scarce.

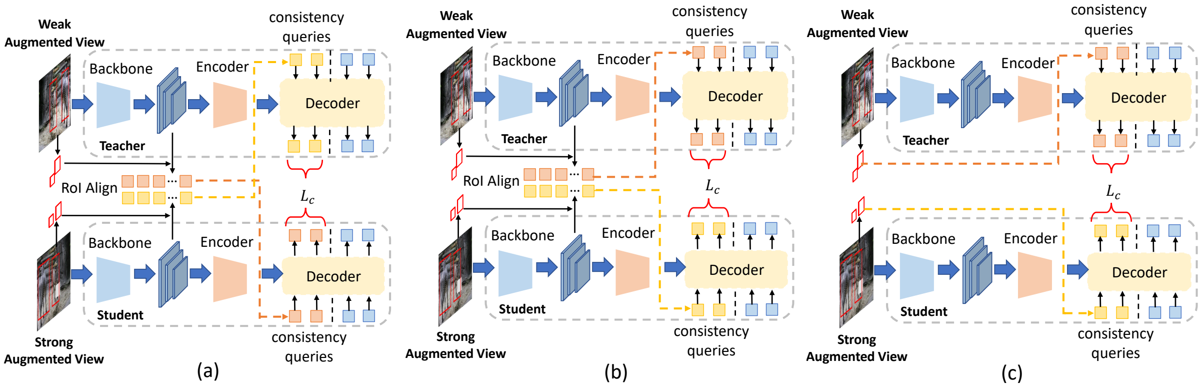


Figure 5. Overview of different variants of the cross-view query consistency. (a) our proposed cross-view query consistency, (b) consistency scheme in Exp-1 which replaces the cross-view consistency queries with single-view consistency queries, and (c) consistency scheme in Exp-2, which directly takes the positional embedding of pseudo boxes as the consistency queries.

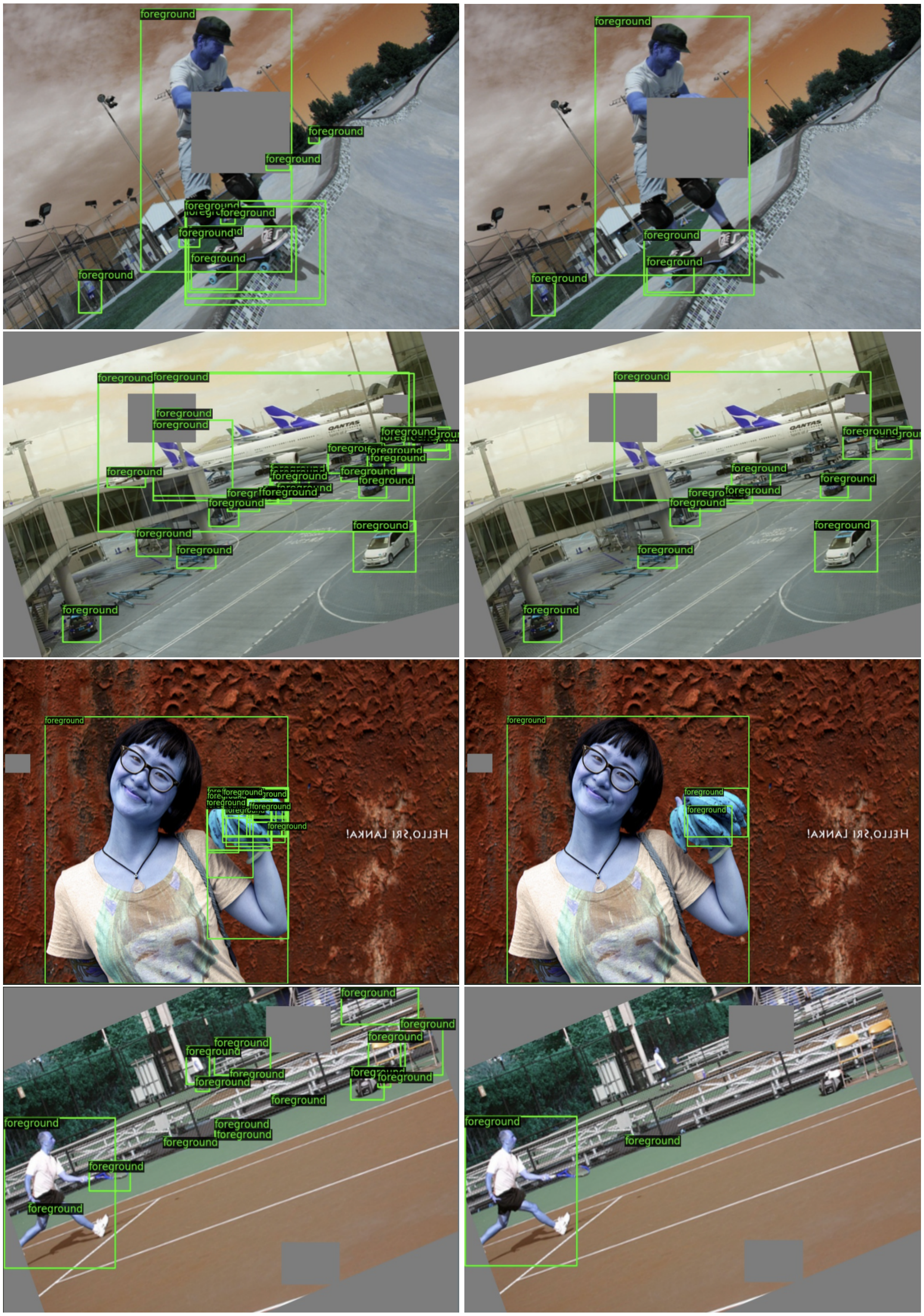


Figure 6. The pseudo boxes before and after the cost-based pseudo label mining. **Left**: the initial pseudo boxes obtained with threshold τ_1, **Right**: the pseudo boxes after the cost-based pseudo label mining with GMM. Note that we applied strong augmentations on the unlabeled images and visualized the predictions accordingly.

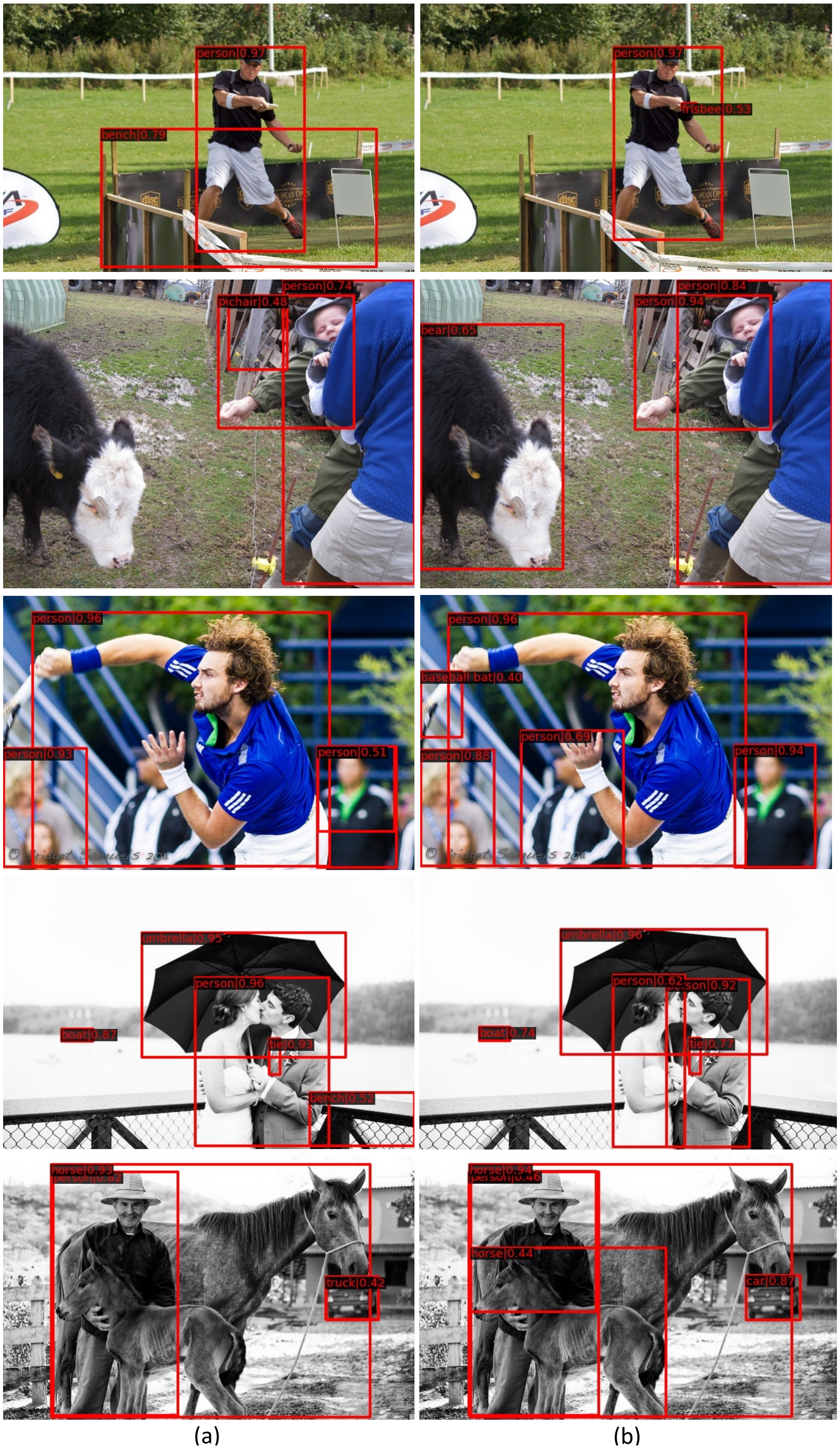


Figure 7. Qualitative comparisons between the pseudo labels generated by (a) training without Stage-wise Hybrid Matching and (b) training with Stage-wise Hybrid Matching. The pseudo label threshold $\tau_s = 0.4$.