ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

Dual Subnet Label Assignment for End-to-End Fully Convolutional Pedestrian Detection in Crowd Scenes

Jing Wang[1,a], Xiangqian Li[1,b], Huazhu Xue[2,c], Zhanqiang Huo[1,d,*]

[1]*School of Software, Henan Polytechnic University, Jiaozuo, China*
[2]*School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo, China*
[a]*wjasmine@hpu.edu.cn,* [b]*3276520242@qq.com,* [c]*xhz@hpu.edu.cn,* [d]*hzq@hpu.edu.cn*
[*]*Corresponding author*

Abstract: *The fully convolutional detector uses a one-to-one (O2O) label assignment strategy removing NMS post-processing operations and realizes end-to-end detection. However, limited positive samples lead to slow convergence of the full convolutional end-to-end detector in crowded scenes. In this paper, we propose a Dual Subnet Label Assignment network (DSLA), which accelerates the model convergence to improve the detection performance and achieves end-to-end detection pipeline. First, we present Soft Label Assignment (SLA), which utilizes soft anchors with positive and negative sample semantics for model learning and accelerates model convergence. Meanwhile, SLA assigns labels to occluded targets in crowded scenes. We combine the SLA branch and the O2O branch for co-training, and achieve end-to-end detection through the O2O branch. Finally, we propose Feature Shuffle to solve the problem of lack of information interaction between different feature layers and highlight the features of the occluded part in a crowded scenes. Experiments demonstrate that DSLA outperforms OneNet in terms of model convergence speed and detection performance. Especially, on the CrowdHuman dataset, our method outperforms SOTA, achieving 91.7 AP, 47.2 MR^{-2}, 80.3 JI, and 98.5 Recall.*

Keywords: *Object Detection, crowd scenes, end-to-end detector, feature fusion*

1. Introduction

Object detection in crowded scenes is a both fundamental and challenging task in the field of computer vision. Previous research works have been rapidly improved with the help of deep neural networks and large datasets. However, most of them require NMS post-processing operations to remove redundant anchors. Especially in crowded scenes, the presence of NMS leads to situations where the detectors appear to have suboptimal selection of occluded targets.

Removing the detector's NMS to achieve end-to-end detection is a current direction for researchers. DETR[1] is a fully end-to-end detection framework that achieves competitive detection performance by removing NMS through the introduction of learnable queries[2] to represent targets. Inspired by DETR, researchers have implemented fully convolutional end-to-end detection on OneNet[3] by using a one-to-one label assignment strategy. OneNet shows that the key to implementing end-to-end detection on fully convolutional networks is to incorporate classification loss in the one-to-one label assignment strategy. By introducing classification loss, assigning only one prediction (positive sample) for ground-truth can effectively remove redundant detection frames, thus removing NMS to achieve end-to-end detection.

However, due to the limited number of positive samples, the use of one-to-one label assignment strategy for fully convolutional detectors reduces the efficiency of detector learning and leads to the problem of slow model convergence. In DETR, this problem is addressed by introducing additional queries. However, the fully convolutional detector does not use the self-attention mechanism and cross-attention mechanism in DETR to avoid repeated predictions.

Researches[4-7] show that the reason for the slow convergence of the fully convolutional end-to-end detector is that the one-to-one label assignment strategy provides the model with limited learnable supervisory signals. Specifically, the one-to-one label assignment strategy only selects a fixed one of the samples for the target to match and ignores other samples which have similar speech information to it. The above findings motivate us to use these similar samples for fully convolutional end-to-end detector to learn. In this paper, we introduce a new soft label assignment approach as a branch to provide more learning signals for the detector. In this branch, we will feed samples that cannot be ignored into the

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

model as learnable supervised signals. The details are shown in Figure1.The soft label assignment branch is mainly used to learn these additional supervised signals to make the detector more focused on important features. We jointly train the Soft Label Assignment branch and one-to-one label assignment branch Once Dual Subnet Label Assignment is done with the training, we keep only one-to-one label assignment branch for end-to-end detection.

Experimental results suggest that our Dual Subnet Label Assignment accelerates the convergence of the fully convolutional end-to-end detector and also improves the detection performance. After 36 epochs of training, our method can surpass the performance of mainstream NMS. detectors. We also introduce the Feature Shuffle feature fusion module for increasing the interactions between detection points between different feature layers to improve the detection performance of the detector in crowded scenes.

Our main contributions are summarized as:

- We propose the Soft Label Assignment strategy, which introduces soft anchors to provide more learning signals for the model, as a way to alleviate the problem of missed and wrong detections in crowd scenes
- We propose Feature Shuffle to increase the information interaction between different feature layers.
- We propose Dual Subnet Label Assignment to accelerate the model convergence by providing more feature learning signals from the Soft Label Assignment branch, and realize end-to-end detection through the one-to-one label assignment branch.
- Compared with the current mainstream fully convolutional end-to-end detectors and anchor-based detectors, our method exhibits better detection performance.

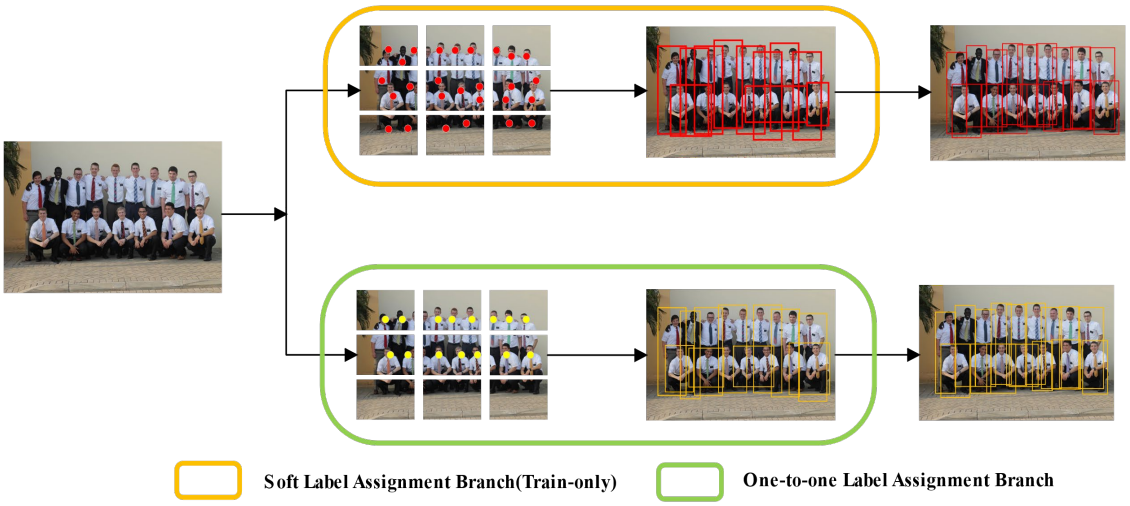


Figure 1: We jointly train the Soft Label Assignment branch and the one-to-one label assignment branch.

2. Related Works

2.1. Fully convolutional detector

Fully convolutional detectors can be further subdivided into one-stage[8-10] detectors[11] and two-stage detectors[12-14] . The main difference between these two types of detectors is that the two-stage detector will have the stage of generating candidate regions, and then the model further removes the repetitions based on the generated candidate regions to finally output the results, while the one-stage detector directly predicts the dense anchor category and location region on the convolutional feature map. Then the shape and size of the anchors should be different for different datasets. To overcome this problem, anchor-free detectors simplify the detection pipeline. FCOS[8] and CenterNet[15] can directly select anchor points as regression target objects.

There are two main reasons for the failure of crowded scene detection: (1) highly overlapping targets may have very similar characteristics, so it is difficult for the detector to generate differentiated predictions for each positive sample individually, and (2) instances may overlap heavily with each other, so the predictions are likely to be incorrectly removed by the NMS. To address the above problems,

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

previous works try to tackle them from different perspectives, e.g., using soft NMS[16] to go out duplicates, new loss functions, etc. However, these methods are either too complex or less efficient in dealing with highly overlapping instances. In order to improve the shortcomings of previous work, MIP proposes multi-instance prediction methods, EMD loss and RM modules to deal with potentially erroneous predictions. Although the new multi-instance prediction and refinement modules can improve the detection performance of the detector, the performance of the detector is still limited due to the presence of post-processing operations such as NMS.

2.2. Query-based detector

As a pioneer of query-based detectors, DETR selects an alternate set of learnable object queries as candidate regions for image feature interaction. It achieves end-to-end target detection with the help of bipartite graph matching and a global attention mechanism. However, DETR suffers from slow proficiency and poor performance in small target detection. Much current research aims to improve the interaction mechanism between the feature map and query in order to obtain more relevant and accurate features to improve the detection performance of small targets. Recent work has found that the limited number of positive samples affects the convergence speed of DETR. Therefore, Group DETR[17] introduces additional positive samples in training to accelerate model convergence. However, in dense scenarios, this introduction of extra positive samples[18] incurs significant computational cost and longer training time. Therefore, in this paper, we make it easier to train end-to-end networks by introducing a class of samples with similar contextual semantics.

In order to solve the problem of missed target detection in crowded scenarios, PDETR[19] proposes the use of sense queries in crowded scenarios. to address the problem of high computational effort of sense queries, PDETR designs local-attention strategy and rectified-attention for self-attention and cross-attention in the decoder, respectively. attention strategy and rectified attention field proximity point set selection strategy, respectively. To alleviate the partial occlusion problem, PDETR proposes a V-Match supervision method. In order to speed up the KM process of computing sense queries with ground-truth, a Fast-Match method is proposed.

3. Proposed Method

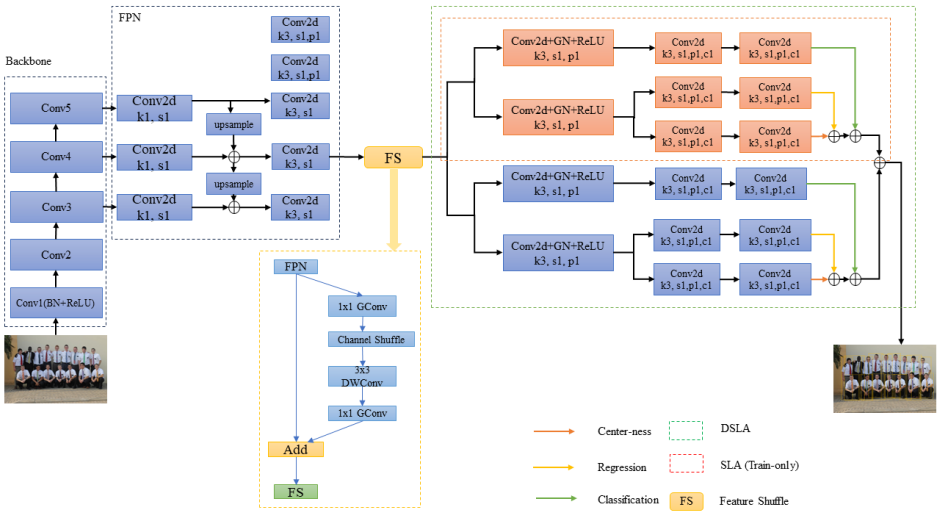


Figure 2: Dual Subnet Label Assignment network pipeline. The green box is Dual Subnet Label Assignment, which contains Soft Label Assignment branch and one-to-one label assignment branch. The red box is the Soft Label Assignment branch, which only participates in the training phase. The structure in the orange box is Feature Shuffle, which is used to enhance the information interaction between different feature layers.

In order to address the drawback of slow model convergence due to the lack of sufficient supervised signals (i.e., enough positive samples) for a fully convolutional detector using a one-to-one label assignment approach, we introduce Dual Subnet Label Assignment, as shown in Figure. 2. Like current fully convolutional end-to-end target detector, our DSLA takes images as input and extracts multi-scale features for classification and regression. Our overall network structure contains backbone, FPN, a

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

feature fusion module, and two subnets. Our approach constructs two branches (Soft Label Assignment branch and One-to-One Label Assignment branch) after feature fusion during training. As shown in the structure of the two-subnet network in Figure. 2, the structure in the green box is the two-subnet label assignment algorithm. In the orange box is the Soft Label Assignment branch, which is only involved in the training phase. Only one branch, One-to-One Assignment, is kept for end-to-end detection during inference.

3.1. Dual Subnet Label Assignment

Our approach is called Dual Subnet Label Assignment. We add Soft Label Assignment approach and one-to-one label assignment approach as two branches to the end-to-end detection network. The Soft Label Assignment branch is only involved in the training phase. These two branches take the shared classification features and regression features as inputs for prediction. Then, the label assignment strategy will be constructed in pairs (ground-truth, Prediction) to compute the loss. Optimizing the DSLA network allocation strategy is a multi-objective optimization problem. In the ideal case, we would like to find the solution where both branch losses are minimized. However, it is almost impossible for a solution to satisfy both losses functions under the same condition. Pareto optimization is a common solution to this problem.

We set different weights for the two branches to ensure that the network is trained to introduce more supervised signals as well as to reduce the computational effort of removing duplicate frames.

$$Loss = \lambda_{o2o} \times L_{o2o} + \lambda_{SLA} \times L_{SLA} \tag{1}$$

where L_{o2o} and L_{SLA} are the loss functions for one-to-one and Soft Label Assignment branches, respectively. λ_{o2o} and λ_{SLA} are the weights of L_{o2o} and L_{SLA} respectively. (A relatively larger weight than the others indicates that the corresponding objective function is more important than the others.) For example, under the inference phase, when L_{SLA} is 0, it indicates that the network does not carry Soft Label Assignment branch and only performs one-to-one label assignment.

Our DSLA combines the advantages of one-to-one label assignment strategy and Soft Label Assignment strategy. Our approach has both a Soft Label Assignment branch that provides enough learning signals to accelerate the convergence of the model and a one-to-one label assignment strategy branch that enables end-to-end. Our approach trains both branches (the one-to-one branch and the Soft Label Assignment branch) simultaneously in the training phase. At the end of training, only the one-to-one label assignment branch is kept. The Soft Label Assignment branch we introduce is only used in the training phase, which is more similar to a plug-in that improves the detection performance with negligible increase in computational resource consumption. The whole network has no other post-processing operations in the evaluation phase, and still belongs to the end-to-end detector.

3.2. Soft label assignment branch

Disadvantages of using one-to-one label assignment strategy in full convolutional detection networks for end-to-end dense detection Due to the limited number of positive samples, it results in a one-to-one label assignment strategy that reduces the feature learning efficiency and ultimately affects the performance of detector. Additional positive samples are introduced in query-based end-to-end detectors to alleviate this problem, but attention operations in Transformer limit application to fully convolutional end-to-end detector.

In this paper, we propose a simple and effective Soft Label Assignment strategy for dense detection. In addition to defining a certain anchor for each target, several soft anchors are defined that can be learned by the model. Introducing such soft anchors in a fully convolutional end-to-end detection network allows the detector to learn more supervised signals. The weight of such model-learning soft anchors is dynamically adjusted during training to allow them to contribute more to representation learning in the early training phase and more to repetitive prediction removal in the later phase.

Soft anchor. Under one-to-one label assignment, the model selects only one anchor as the positive sample and ignores the remaining soft samples that can provide learning signals for the model. Especially in crowded scenes, these supervised signals with learnable signals cannot be ignored during the training process.

As shown in Figure 3, the one-to-one label assignment approach uses only the red box part as a supervised signal. In terms of semantic information, the blue box and the red box do not differ much, and

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

both can provide learnable supervisory signals for the model. The Soft Label Assignment strategy selects one feature's anchor as a completely certain anchor (the red box), and selects several soft anchors to give them positive and negative weights to feed into the model. The positive and negative weights of the soft anchors are dynamically adjusted during the training process so that the network can learn a strong feature representation and achieve end-to-end detection capability at the same time.

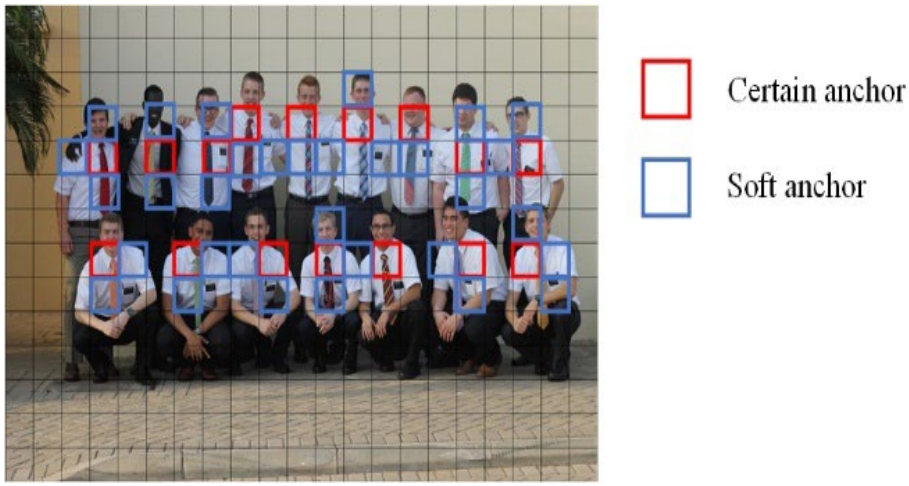


Figure 3: The comparison between one-to-one label assignment and soft label assignment. The certain anchor in the red box has similar semantic information to the soft anchor in the blue box.

3.3. Select certain positive anchor

In the Soft Label Assignment strategy, a specific positive anchor will be selected for each instance. Previous end-to-end detectors have used a selection metric for a predictive-aware mechanism, which takes into account the costs of classification and regression to select a uniquely positive sample. Following this principle, both the classification score and IoU are incorporated into the selection metric for a particular anchor, which is defined as:

$$S_{i,j} = \mathbb{F}\left[i \in \Omega_j\right] \times p_{i,cj} \times IoU\left(b_i, b_j\right) \tag{2}$$

where $S_{i,j}$ denotes the score of the match between anchor i and instance j. The spatial prior is used in the o2o and SLA approaches. $P_{i,cj}$, b_j This spatial prior is commonly used in o2o and SLA methods because the observed anchor in the center region of an instance is more likely to be a positive anchor.

The anchors can be sorted in descending order based on the metric $S_{i,j}$. Prior work has typically formulated forward-anchor selection as a bisection matching problem and solved it by the Hungarian algorithm. In order to reduce the computational effort, in the forward-anchor selection part, we directly select the one with the highest score as the specific forward-anchor for each instance.

3.4. Assigning labels to soft anchors

In addition to specific positive anchors, based on the score $S_{i,j}$ selects anchors as soft anchors because they have similar semantic context as specific positive anchors. Dynamic soft labels are assigned to these soft anchors in order to reduce the possibility of repeated predictions. Assuming that there are N Epochs to train the network, the classification loss of each soft anchor i during the j epoch training period is defined as:

$$t_i^j = \frac{P_i}{\max kp_k} \tag{3}$$

$$T^j = \frac{T^{\min} - T^{\max}}{N-1} \times j + T^{\max} \tag{4}$$

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

where P_i is the predicted categorization score of anchor i, and $\left(1-t_i^j\right)$ are the governmental degree of this anchor at the j Epoch, respectively.

where T_j is the time-dependent variable where is assigned the same value for all samples in the j epoch, and $T^{\max}$ and $T^{\min}$ control the degree of positivity of the soft anchor for the first epoch and the last epoch, respectively. Setting the loss weights to be positively correlated with the assigned scores takes into account the fact that an anchor with a higher prediction score should provide more supervised signal for positive sample signals. Using P^i directly as a weight will destabilize the training of difficult samples because their prediction scores are much smaller than the prediction scores of simple samples. Therefore, use the ratio between P^i and max to normalize the weights of different samples to the same scale.

Dynamic tuning of T_j is important because it controls the trade-off between feature learning and elimination of duplicates at different stages of the training phase. In the early training phase, T_j is set relatively large to introduce more positively supervised signals for representation learning so that the network can quickly converge to a robust feature representation space. As training progresses, the weight of the soft anchor is gradually reduced so that the network can learn to remove duplicate predictions.

For each instance, a specific positive Anchor and a soft anchor are selected. The remaining Anchor is set to negative samples, and the training target for the classification branch is specified for each instance:

$$L_{cls} = BCE\left(p_c, 1\right) + \sum_{i \in A} BCE\left(p_i, t_i\right) + \sum_{i \in B}\left(p_i, 0\right) \tag{5}$$

where pc is the classification score of a single specific anchor, A and B denote the set of soft anchor and negative sample anchor respectively. BCE denotes Binary Cross Entropy Loss and FL denotes Focal Loss. regression loss is defined as:

$$L_{reg} = \sum\nolimits_{i \notin B} GIoU\left(b_i, b_{gt}\right) \tag{6}$$

Where the $GIoU$ loss is based on the general intersection loss on the concatenated set, bi is the predicted position of anchor i, and is the position of the GT corresponding to anchor i. Note that the regression loss is applied to both certain anchor and soft anchor.

3.5. Assigning labels to soft anchors

In Anchor-free detectors, anchor points act like query in query-based detectors. However, there is no way for the attention mechanism and query to be able to interact with each other in a fully convolutional network as in Transformer. This leads to a congested scenario, which leads to an allocation failure in label assignment during detection. Inspired by DETR, we treat the anchor points in the FCOS network as query and increase the interaction between the anchor points by introducing the Feature Shuffle module.

Inspired by the feature interactions between different channels in ShuffleNet[20] and RelationNet[21], we propose Feature Shuffle to compensate and increase the interactions between different levels of anchor points, as shown in Figure. 2. Where S channels across S neighboring levels are shuffled to form a new feature pyramid. In detail, the feature map of level i exchanges S channels with the feature maps of levels $i-1$ and $i+1$ at the same time. Considering the different spatial dimensions on the feature pyramid, a bilinear interpolation method is used for the exchange. We apply Feature Shuffle after FPN, which allows the interaction of anchor points at different scales. This approach stabilizes the training and improves the detection performance, while the additional computational cost is negligible.

4. Experiment and Analysis of Results

4.1. Datasets

In this section, we evaluate our method on the CrowdHuman dataset and the COCO dataset. Outperforming our proposed method mainly to improve the detection in crowded scenes, we perform most of the comparison experiments and ablation experiments on the CrowdHuman dataset. Experiments

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

on the COCO dataset verify that our method improves detection performance even in generalized scenarios.

The CrowdHuman dataset contains training set, validation set and test set containing 15000, 4370 and 5000 images respectively. On average, there are about 23 individuals per image in the CrowdHuman dataset, as shown in Table 1.The Crowd Human dataset has rich scenarios, different pedestrian poses, and different levels of crowding. Also, CrowdHuman dataset provides three different bounding box annotations for each pedestrian instance, namely head bounding box, visible area bounding box, and full body bounding box.

The COCO dataset is very large, with 118,287, 5,000, and 40,670 images for the training, validation, and test sets, respectively, and contains multiple target categories and detailed annotations. Each image in the COCO dataset contains approximately nine targets, as shown in Table 1. The COCO dataset is annotated for each target with the location of the category and bounding box.

Table 1: Instance densities for each dataset. The threshold for overlap statistics is $IoU > 0.5$. **Averaged over the number of categories.*

Dataset	#object/img	# overlaps/img
CrowdHuman[31]	22.64	2.40
COCO[32]	9.34	0.0015

4.2. Evaluation metric

Evaluation of Indicators. We mainly use four important metrics, AP, MR^{-2}, JI and Recall.

Averaged Precision (AP). This is the most commonly used metric in the field of object detection. AP reflects the accuracy of the detection results. A larger AP indicates a better performance of detector.

MR^{-2} is the logarithmic mean miss rate of false positives per image in $[10^{-2},100]$. This metric is commonly used in pedestrian detection.MR^{-2} is very sensitive to false positives, especially false positives with high confidence will significantly impair the MR^{-2} rate. A smaller MR^{-2} indicates better performance.

Jaccard Index (JI) is primarily used to assess the counting ability of a detector. Unlike AP and MR^{-2} defined on a predicted sequence with decreasing confidence, JI assesses the degree of overlap between the predicted set and the true frame. Typically, prediction sets can be generated by introducing a confidence score threshold. In this paper, for each evaluation entry, we compute the optimal JI score by exploring all possible no center thresholds. Larger JI indicates better performance.

Recall. Recall is the fraction of samples that the classifier considers to be positive and are indeed positive as a proportion of the samples shown to be positive. Higher recall indicates better performance.

4.3. Implementation details

We chose to use OneNet as baseline and also as O2O branch for end-to-end detection. We use the pre-trained ResNet50[26] on ImageNet[27] as backbone and choose AdamW[28] as our optimizer. Meanwhile, our learning rate is defaulted to 0.001. Other hyper-parameters settings, such as batch size and weight decay, are consistent with OneNet. We train with 36 epochs training schedule. In terms of experimental environment, we conduct our experiments based on PyTorch 1.9.0[29] and MMdetection[30]. In terms of hardware, we use four 40G Tesla A100s for training and testing.

4.4. Results on the CrowdHuman dataset

In order to holistically and comprehensively evaluate the method proposed in this paper, we conducted extensive experiments on the CrowdHuman dataset. The evaluation is based on four evaluation metrics, with AP as the main evaluation metric. In Table 2, we give the experimental results of the method proposed in this paper as well as the results of the current mainstream detectors. For a fair comparison, all the anchor-free detectors, anchor-based detectors and baselines in the table use ResNet50 pre-trained on ImageNet as the backbone. We compare the current mainstream detectors including Anchor-based and Anchor-free detectors. Detectors.

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

Table 2: Results for different detectors on the CrowdHuman validation set. All detectors use ResNet50 pre-trained on ImageNet as the backbone. +MIP denotes multi-instance prediction using the set NMS as post-processing.

Method	*AP*(%)	*MP*$^{-2}$	*JI*(%)	*Recall*(%)
RetinaNet[9]	85.3	55.1	73.7	-
ATSS[22]	87.0	51.1	75.9	95.9
ATSS[22]+MIP[11]	88.7	51.6	77.0	-
FPN[23]+NMS	85.8	42.9	79.8	-
FPN[23]+soft NMS[16]	88.2	42.9	79.8	-
NGLA[24]	89.5	46.6	-	96.2
FPN+MIP[11]	90.7	41.4	82.4	-
FCOS[9]	86.6	54.0	75.7	-
FCOS[8]+MIP[11]	87.3	51.2	77.3	-
POTO[25]	89.1	47.8	79.3	97.9
FCOS[8]+NGLA[24]	90.1	45.6	-	96.6
baseline[3]	90.1	50.0	78.2	97.9
Ours	**91.7**	**47.2**	**80.3**	**98.5**

As demonstrated by the results in Table 2, our method outperforms these well-established detectors and achieves a significant performance improvement compared to the Anchor-based and Anchor-free detectors. This indicates that our method can effectively address the problem of pedestrian occlusion in crowded scenes. In particular, our method achieves 1.6 higher AP than OneNet as well as 2.8 and 2.1 improvement in MR^{-2} and JI, respectively. Especially, we show SOTA in Recall.

As shown in the left part of Figure.4 (a), our method converges faster with the same training parameter settings compared to OneNet. At the same time, our method exhibits better performance with the same epochs settings.

Compared with OneNet, our method converges faster with the same training parameter settings, as shown in Figure. 4(a). The convergence curve of the model of our method stabilizes at the tenth epoch, while OneNet starts to stabilize only at the fourteenth epoch. Meanwhile, the detection performance of our method performs better.

Table 3: Results of current mainstream end-to-end detectors on CrowdHuman validation set

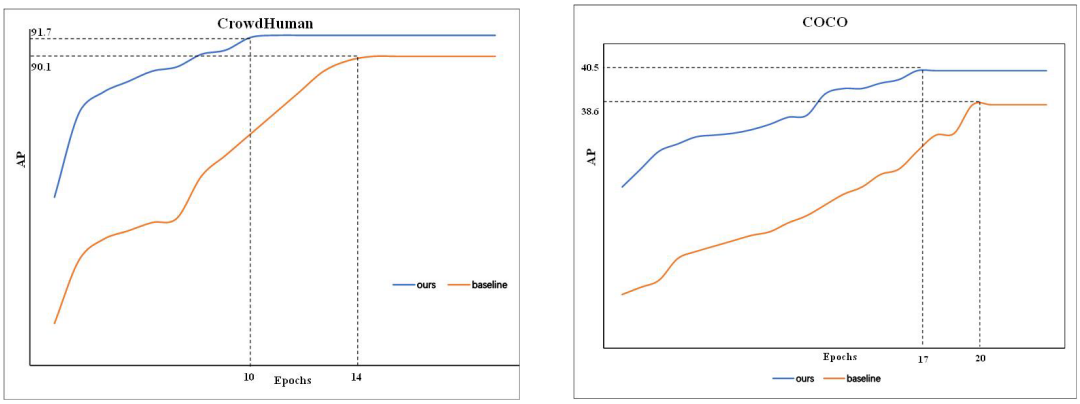
Method	AP(%)	MR^{-2}	Recall(%)
DETR[1]	75.9	73.3	-
DeFCN[25]	89.1	47.8	97.9
PDETR[19]	91.6	43.7	-
D-DETR[4]	91.5	43.7	-
baseline[3]	90.1	50.0	-
Ours	91.7	47.2	98.5

The current mainstream end-to-end detectors are the DETR family and full convolutional networks, respectively. As shown in Table 3, our method outperforms current query-based detectors. Compared with PDETR, our method also has a small lead. In terms of the fully convolutional end-to-end detector, our method improves 2.6 AP and 0.6 MR^{-2} compared to DeFCN[25]. In terms of Recall, we have a significant improvement, which suggests that the soft anchor and the certain anchor we introduce in the SLA branch have the same semantic information and can be used for the model to learn features simultaneously.

The visualization results of DSLA on the CrowdHuman dataset are shown in Figure. 5. The

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

experimental and visualization results show that our method can effectively mitigate the problem of inaccurate detection and localization due to occlusion in crowd scenes. Our method is able to show competitive performance in different occlusion situations, such as pedestrian-to-pedestrian occlusion and pedestrian-to-object occlusion.



(a) Convergence curve of DSLA on CrowdHuman *(b) Convergence curve of DSLA on COCO*

Figure 4: (a) is the convergence curve of DSLA on CrowdHuman, (b) is the convergence curve of DSLA on COCO dataset. Both DSLA and baseline use ResNet50 as backbone, and both complete training under the same 36 epochs. Our experiments are done under PyTorch and MMdetection

Figure 5: Visualized detection results on CrowdHuman, where the method proposed in this paper is able to present better detection results for pedestrians in different backgrounds, crowdedness, occlusion between pedestrians and pedestrians, occlusion between pedestrians and objects, and detection of pedestrians in different poses. The first row is dense occlusion between pedestrians; the second row is occlusion between pedestrians and objects; the third row is occlusion of pedestrians in both near and far views; and the fourth row is sparse occlusion.

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

Ablation experiments on CrowdHuman. The ablation experiments were performed to verify the effectiveness of the SLA branching and Feature Shuffle feature fusion methods. Table 4 shows that after adding SLA branch, AP, MR^{-2}, and Recall have 1.0, 2.2, and 0.3 enhancements, respectively. After the addition of FS, although there is an improvement in all indicators, the improvement is very small. From the results in Table 4, it can be seen that SLA can effectively solve the problem of pedestrian occlusion in congested scenarios. At the same time, FS also enables SLA to show can good performance, indicating its necessity and effectiveness.

Table 4: Ablation experiments on CrowdHuman

Model	FS	DSLA	AP	MR^{-2}	Recall
baseline[3]	✗	✗	90.1	50.0	97.9
ours	✓	✗	90.5	49.6	97.5
ours	✗	✓	91.1	47.8	98.2
ours	✓	✓	91.7	47.2	98.5

In this paper, the original intention of our proposed method is to be able to accelerate the end-to-end detector convergence speed without increasing the computational resources. Combing the results in Table 5 and the graph in Figure. 4(a), our method only adds negligible computation in the model training phase. In the evaluation phase, our computation remains consistent with OneNet. The results show that our method not only effectively solves the problem of slow convergence speed of the end-to-end fully convolutional detector, but also improves the performance of the detector without consuming additional computational resources.

Table 5: Ablation experiments on CrowdHuman

Model	phase	Params	FLOPs
baseline[3]	infer	32.0M	201
ours	infer	32.0M	201
ours	train	32.2M	204

4.5. Influence of different hyper-parameter settings in SLA

In the SLA branch, we want the number of generated SOFT anchors to provide a sufficiently large number of learnable signals for the model. At the same time, that number of samples does not increase the additional computational resources generated by removing duplicate frames due to redundancy.

Table 6: Ablation experiments on CrowdHuman numbers of soft anchors

K	10	9	8	7	6	5	4
AP	90.2	90.7	91.2	91.2	**91.7**	90.8	90.0
Recall	96.5	96.8	97.2	97.8	**97.8**	97.5	98.6

As shown in Table 6, when we set the number of soft anchors to 6, the best detection performance is achieved on the CrowdHuman val set without consuming additional computational resources. So in the following experiments, we all set the soft anchor quantity to 6 by default.

Table 7: Experimental results of close-up retrieval

	T^{max}				
T^{min}	*0.4*	90.8	91.7	89.2	88.7
	0.3	90.5	90.4	90.0	88.6
	0.2	90.3	90.4	89.4	88.2
	0.1	89.0	90.1	89.9	88.6

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

In the SLA branch, we want the soft anchor to be able to occupy a large weight in the initial stage of training, while in the end stage of training, we want to reduce the dependence on the SLA branch and focus on the O2O branch. The two parameters $T^{\max}$ and $T^{\min}$ control the positive and negative weights of the soft anchor in the first and the last Epochs, respectively, during the training process. As shown in Table 7, different results were obtained by adjusting different sizes of $T^{\max}$ and $T^{\min}$. It can be seen that when $T^{\max}$ is taken as 0.6, a reduced $T^{\min}$ will decrease the detector performance, which suggests that a large $T^{\min}$ is also important to increase the performance of the detector in terms of the learning provided in the final training phase. When $T^{\min}$ is taken as 0.4, increasing or decreasing $T^{\max}$ will affect the performance of the detector, which suggests that soft anchors with similar semantics affect the performance of the model in the de-duplication phase during training.

4.6. Results on the COCO dataset

In this paper, our proposed DSLA is used to solve the problem of downstream earthly masking in congested scenarios. There are also some mildly congested scenarios in the COCO dataset. We do experiments and ablation experiments on the COCO dataset to evaluate our approach.

Our experiments are scheduled on MMdetection. Backbone uses a pre-trained ResNet50 model on the ImageNet dataset. The optimizer uses AdamW with a default learning rate of 4e-4. The batch size and weight decay settings are kept the same in order to compare the fairness of the experiments. The results in Table I are all obtained after 36 epochs of training.

As shown in Table 8, our method exhibits good performance on the COCO dataset. With the same parameter settings, our method has a 1.8 AP improvement over OneNet. The improvement is 1.9 and 2.0 on AP50 and AP75 results, respectively. The results demonstrate that DSLA also shows good performance for mildly congested and multi-category target detection datasets. The visualization results of DSLA on the CrowdHuman dataset are shown in Figure. 7.

As shown by the graph in Figure. 4(b), when our method is trained on the COCO dataset, the model converges at the seventeenth epoch, whereas OneNet starts at the twentieth epoch before the model gradually begins to converge. Experiments demonstrate that our approach accelerates model convergence on the generalized target detection dataset as well, and also improves detector performance.

Table 8: Experimental results of close-up retrieval

Model	*AP*(%)	AP_{50}(%)	AP_{75}(%)
FCOS[8]	36.7	56.8	39.6
RetinaNet[9]	37.2	56.7	39.0
baseline[3]	38.6	56.6	42.2
Ours + FS	39.1	57.2	42.7
Ours + DSLA	39.5	58.1	43.8
Ours+DSLA+FS	40.5(+1.8)	58.7(+1.9)	44.2(+2.0)

4.7. Visual analysis of Feature Shuffle

Based on the comparison between the upper and lower parts of Figure. 6, it can be observed that the features after Feature Shuffle not only reduce the redundant parts, but also strengthen the network's focus on the features. Especially in the case where pedestrians are occluded, after Feature Shuffle, the image features mainly focus on the pedestrian aspect. Especially when pedestrians are occluded by objects, after Feature Shuffle, the image features of characters are more obvious.

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

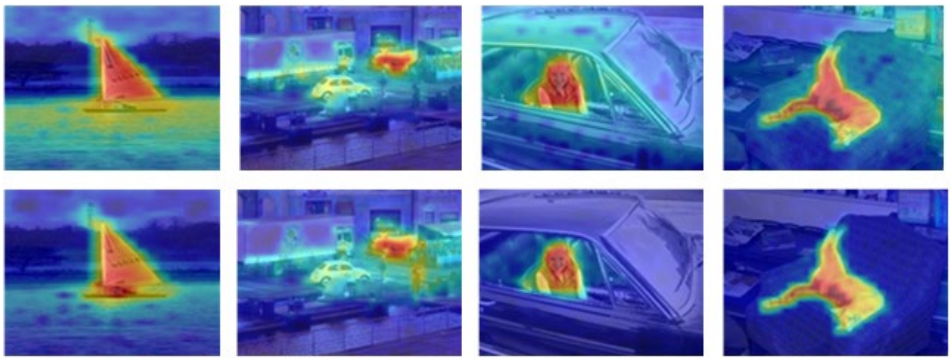


Figure 6: Visualization of score map. We use the heat map from to display the features. The upper part is the feature map of first-stage of ResNet50. The lower part is the feature map after Feature Shuffle. The visualization shows that FS is able to generate representative features.

Figure 7: Visualized detection results on the COCO dataset. The first and second rows are mildly crowded scenes, and pedestrians show different postures, and pedestrians and objects are also closely connected, and our method is able to detect them well for different targets as well. The third, fourth and fifth rows gradually change from light to heavy congestion, with occlusion between pedestrians and between pedestrians and object.

5. Conclusions

In this paper, we propose Dual Subnet Label Assignment (DSLA) to solve the problem of slow convergence of the model. The SLA branch provides the soft anchor to the model so that the model can learn more learning signals. DSLA combines the advantages of the SLA branch and the O2O branch to accelerate the model convergence and improve the detection performance without bringing extra computational resources. DSLA combines the advantages of SLA branch and O2O branch, which can accelerate the model convergence and improve the detection performance without bringing extra computational resources. Feature Shuffle strengthens the interactions between different feature layers, so that the detector can focus more on the target features. Through experiments, we demonstrate that our approach not only shows good performance in crowded scenarios, but also improves the performance of the detector in generalized scenarios. Further in the future, we hope that DSLA can be adapted to as many end-to-end detectors as possible. In addition, how to dynamically adjust the number of soft anchors in

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

SLA branches for different degrees of occlusion is also our future research direction.

References

[1] CARION N, MASSA F, SYNNAEVE G. End-to-end object detection with transformers; proceedings of the European conference on computer vision, pp. 213-229, 2020.
[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[3] Z hang S, Wang X, Wang J, et al. What are expected queries in end-to-end object detection?[J]. arXiv preprint arXiv:2206.01232, 2022.
[4] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv:2010.04159, 2020.
[5] Chen Y, Zhang Z, Cao Y, et al. Reppoints v2: Verification meets regression for object detection[J]. Advances in Neural Information Processing Systems, 2020, 33: 5621-5631.
[6] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.
[7] Z hu B, Wang J, Jiang Z, et al. Autoassign: Differentiable label assignment for dense object detection[J]. arXiv preprint arXiv:2007.03496, 2020.
[8] Tian Z, Shen C, Chen H, et al. FCOS: Fully convolutional one-stage object detection [J]. arXiv preprint arXiv:1904.01355, 1904.
[9] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
[10] FENG C, ZHONG Y, GAO Y. Tood: Task-aligned one-stage object detection; proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, pp: 3490-3499, 2021.
[11] CHU X, ZHENG A, ZHANG X. Detection in crowded scenes: One proposal, multiple predictions; [C]//Proceedings of the IEEE international conference on computer vision, pp: 12214-12223, 2020.
[12] WANG J, ZHAO C, HUO Z. High quality proposal feature generation for crowded pedestrian detection. Pattern Recognition vol.128, pp.108605 2022.
[13] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
[14] Li X, Wang W, Wu L, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[J]. Advances in Neural Information Processing Systems, 2020, 33: 21002-21012
[15] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6569-6578.
[16] Bodla N, Singh B, Chellappa R, et al. Soft-NMS--improving object detection with one line of code[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5561-5569.
[17] Chen Q, Chen X, Zeng G, et al. Group detr: Fast training convergence with decoupled one-to-many label assignment[J]. arXiv preprint arXiv:2207.13085, 2022, 2(3): 12.
[18] Sun P, Zhang R, Jiang Y, et al. Sparse r-cnn: End-to-end object detection with learnable proposals[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14454-14463.
[19] Lin M, Li C, Bu X, et al. Detr for crowd pedestrian detection[J]. arXiv preprint arXiv:2012.06785, 2020.
[20] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848-6856.
[21] HU H, GU J, ZHANG Z, Relation networks for object detection; [C]//Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3588-3597, 2018.
[22] ZHANG S, CHI C, YAO Y. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection; [C]//Proceedings of the IEEE conference on computer vision and pattern recognition, pp.6579-9568, 2020.
[23] LIN T-Y, DOLLáR P, GIRSHICK R. Feature pyramid networks for object detection; [C]//Proceedings of the IEEE conference on computer vision and pattern recognition, pp.2117-2125, 2017.
[24] JIANG H, ZHANG X, XIANG S J I T O M. Non-maximum Suppression Guided Label Assignment for Object Detection in Crowd Scenes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2023.
[25] WANG J, SONG L, LI Z. End-to-end object detection with fully convolutional network;

ISSN 2616-5775 Vol. 7, Issue 5: 1-14, DOI: 10.25236/AJCIS.2024.070501

[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, pp.15846-15858, 2021.
[26] HE K, ZHANG X, REN S. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.
[27] DENG J, DONG W, SOCHER R. Imagenet: A large-scale hierarchical image database; proceedings of the 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255, 2009.
[28] LOSHCHILOV I, HUTTER F J A P A. Decoupled weight decay regularization, IEEE conference on computer vision and pattern recognition, 2017.
[29] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library[J]. Advances in neural information processing systems, 2019, 32.
[30] Chen K, Wang J, Pang J, et al. MMDetection: Open mmlab detection toolbox and benchmark[J]. arXiv preprint arXiv:1906.07155, 2019.
[31] Liu S, Huang D, Wang Y. Adaptive nms: Refining pedestrian detection in a crowd[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6459-6468
[32] LIN T-Y, MAIRE M, BELONGIE S. Microsoft coco: Common objects in context; proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part, vol.13, 2014. Springer.