



HAL
open science

Gaussian Balanced Sampling for End-to-End Pedestrian Detector

Yang Yang, Jun Li, Biao Hou, Bo Ren, Xiaoming Jiang, Jinkai Cheng,
Licheng Jiao

► **To cite this version:**

Yang Yang, Jun Li, Biao Hou, Bo Ren, Xiaoming Jiang, et al.. Gaussian Balanced Sampling for End-to-End Pedestrian Detector. 5th International Conference on Intelligence Science (ICIS), Oct 2022, Xi'an, China. pp.318-325, 10.1007/978-3-031-14903-0_34. hal-04666416

HAL Id: hal-04666416

<https://hal.science/hal-04666416v1>

Submitted on 1 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



This document is the original author manuscript of a paper submitted to an IFIP conference proceedings or other IFIP publication by Springer Nature. As such, there may be some differences in the official published version of the paper. Such differences, if any, are usually due to reformatting during preparation for publication or minor corrections made by the author(s) during final proofreading of the publication manuscript.

Gaussian Balanced Sampling for End-to-End pedestrian detector

Yang Yang¹, Jun Li¹, Biao Hou¹[0000-0002-1996-186X], Bo Ren¹, Xiaoming Jiang¹,
Jinkai Cheng¹, and Licheng Jiao¹[0000-0003-3354-9617]

Xidian University, No. 2 South Taibai Road, Xi'an, Shaanxi, China

Abstract. Recently, NMS-free detector has become a research hotspot to eliminate negative influences, while NMS-based detector mis-suppress objects in crowd scene. However, NMS-free may face the problem of sample imbalance that affects convergence. In this paper, Gaussian distribution is adopted to fit the distribution of the targets so that samples can be chosen according to it. And we propose Gaussian Balance Sampling strategy to balance positive and negative samples actively. Besides, a simple loss function, PDLoss, is proposed to eliminate duplicated matches on the label assignment procedure and increase training speed. In addition, by a novel Non-target Response Suppression method, the designed network can focus more on hard samples and improve model performance. With these techniques, the model achieved a competitive performance on the CrowdHuman dataset.

Keywords: Gaussian Balanced Sampling · End-to-end · NMS-free · Pedestrian Detection · PDLoss.

1 Introduction

In the one-to-one NMS-free network, each target has only one positive sample. These positive samples need to be the most representative points for targets as Wang explored in [14]. In addition, too many negative samples will make the model difficult to focus on positive samples because negative samples will contribute more to gradient backpropagation. Therefore, it is necessary to select an appropriate number of negative samples. For the selection of negative samples, OneNet[12] and DeFCN[14] directly select all the points as negative except the positive ones. Most of the areas that belong to the background will bring too many easy negative samples.

Another noteworthy problem is that when two targets are highly overlapped, they could match the same sample point simultaneously. However, the sample points can only keep one target generally. For the one-to-one assignment model, the situation gets worse due to the other target getting discarded straightly because each target has only one positive sample. Some works like DeFCN[14], adopt Hungarian matching to eliminate duplicated matching. But the some of the positive samples may be non-optimal and the matching algorithm cost lots of time.

To solve these problems, we propose a sample selection method and duplicated matching elimination strategy. And our method has achieved excellent performance on the CrowdHuman[10] dataset.

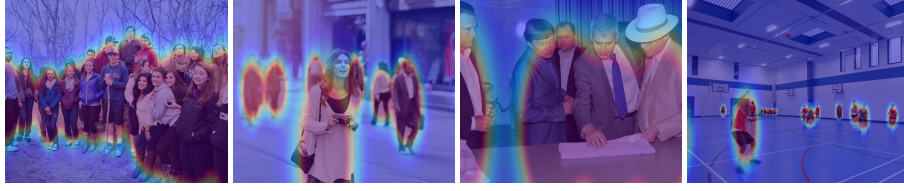


Fig. 1: The mapping of the extracted features onto the original image. The mapping is somewhat smooth. The response value is larger near the target and smaller far away.

2 Approach

2.1 Gaussian Negative Sampling

We found in previous work that in the inference results of the one-to-one matching model, the confidence score is centered on the positive sample point with the highest score, and approximately decays with a Gaussian distribution. Therefore, we assume that the representative strength of the target feature also conforms to this distribution. The feature response of targets gets high in the keypoint and gradually decreases to the outward. Therefore, a two-dimensional Gaussian model is employed to represent the distribution of features:

$$f(x) = \lambda e^{\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right]} \quad (1)$$

$$\lambda = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

where x, y are the coordinate positions. For simplicity, we regard x and y as independent of each other. Then the formula can be simplified to:

$$f(x) = \lambda e^{\left[-\frac{1}{2}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right]} \quad (2)$$

The mean values μ_1, μ_2 of Gaussian distribution are the coordinates of positive sample points x_p, y_p . And the variance values σ_1, σ_2 of Gaussian distribution are set as half the length h and half the width w of the bounding box. We fit the distribution for each target and stack them. Thus, the response of features on the original input image I can be described.

We introduce an additional factor to normalize the values of gaussian distribution for different density of targets. So we set λ to $1/N_o^i$. N_o^i is the number of boxes that are overlapped with the box i .

Since the response graph with all targets is obtained, the positive samples can be chosen as positions with high distribution response. For the selection of negative samples, samples can be directly selected as the area with the response larger than a certain threshold.

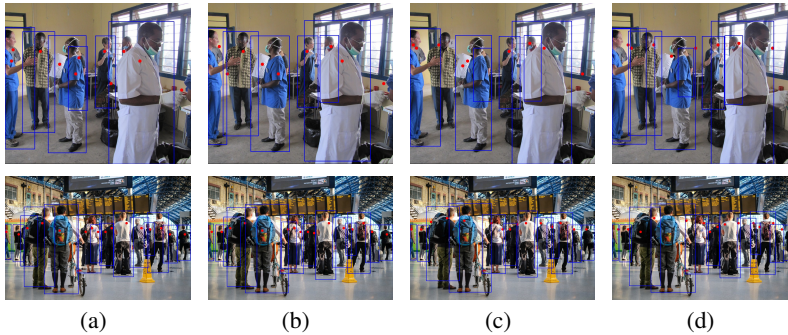


Fig. 2: The matching result of positive sample points and the final detection.

2.2 PD Loss

In the process of label assignment, there is a case that two different targets match the same sample point simultaneously. Each sample point can only match one target, so repeated matching will lead to lose one of the targets. After our verification, in OneNet[12], it actually has the circumstance that two targets match the same sample at the same time, and one of them is discarded directly. This possibility may be relatively small since we only select one positive sample and there are more other samples to choose from, we have no reason to discard it. When the target becomes denser and denser, this risk will further expand, and we cannot ignore it at this time. In addition, we find that the overlapping of sample points will also lead to confusion when learning target features, which will affect the convergence speed.

DeFCN[14] utilizes Hungarian algorithm[5] to eliminate duplication, but its training convergence speed is too slow, which greatly increases the cost. Therefore, we hope to find a more elegant and efficient method. Inspired by Repulsion loss[15], we design a loss function to prevent duplication, called PDLoss:

$$\mathcal{L}_p = \lambda_p \sum_i^n \sum_j^n \hat{dist} \quad (3)$$

where \hat{dist} is the inverse distance between match points:

$$\hat{dist} = \frac{1}{k} \times \frac{\text{Sigmoid}(\mu - |x_i - x_j|)^{|x_i - x_j|}}{|x_i - x_j| + \epsilon} \quad (4)$$

where $|x_i - x_j|$ represents the pixel distance between two matching points. k is a constant set to 0.5. When two targets overlap, the distance is 0, so we need to limit them to $1/k$ to prevent overflow. Without duplicated matching, the loss still exists in a certain range μ . This will make the distance between the close matching sample points as far as possible so as to obtain more discriminative features. ϵ set to prevent formulas from causing calculation errors if the denominator is 0. With a simple loss function, we achieve Hungarian matching with almost no increase in training time. At the same time,

the samples of different targets are kept as far away as possible to avoid false detection caused by the similarity of close-range features.

2.3 Non-target Response Suppression

The non-target response suppression mainly includes two parts, one is about the loss function in training, the other is about the score weighting in inference. In front, we select the sample points according to the Gaussian distribution. Among these sample points, the one near the center is more representative. The negative samples close to the positive sample points are difficult to distinguish and should receive more attention; while the samples that are far away maybe belong to the background and relatively easy to distinguish, so we can pay less attention to them. Therefore, we use Gaussian distribution matrix to weigh the loss of each sample point:

$$\mathcal{L}_w = \hat{f}(x_s)\mathcal{L}_{reg}, \quad x_s \in S \quad (5)$$

where S is all the selected samples. This method is very similar to the attention mechanism.

3 Experiments

Our experiments are implemented by PyTorch on a server with 8 NVIDIA Tesla V100 and 32G memory. Given that NMS-free’s approach is more competitive in dense scenarios, we perform experiments on crowd scene benchmarks to verify the generality of our methods. We trained every model for 30K iterations on Detectron2[16] with 2 images per GPU, which is consistent with the experimental setting of the baseline. For the model, we use ResNet50[4] as our backbone and the initialized weights are pre-trained on ImageNet[2]. The learning rate is initially set to 0.01 and then decreased by 10 at the 20K and 25K iterations. We use SGD to optimize all the models with a weight decay of 0.0001 and a momentum of 0.9.

3.1 Performance on CrowdHuman dataset

Because the NMS-free network is more competitive in dense scenarios, we verify our model on the CrowdHuman[10] dataset. Under the same experimental conditions and ImageNet[2] pre-trained backbone, we compare the performance of our method with the other two NMS-free networks and the traditional methods on the dataset. All the models are trained with the same setting: batch size of 16.

From the table??, we can see that our model can achieve a higher recall than the traditional models. In addition, our method is superior to other one-to-one networks due to its excellent sampling ratio. To be mentioned, OneNet only experiments on the COCO dataset in its paper and this result is obtained by our transfer training on the CrowdHuman dataset. In our experiment, the batch size of OneNet is set to 16, which is different from its reported 64 batch size. The worse performance of OneNet may be caused by this.

Table 1: The Performance of different model on CrowdHuman dataset.

Method	AP	mMR	Recall
RetinaNet[6]	81.7	57.6	88.6
Faster RCNN[8]	84.95	50.49	90.24
Adaptive NMS[7]	84.71	49.73	91.27
PS-RCNN[3]	86.05	-	93.77
FCOS[13]	86.1	55.2	94.3
ATSS[17]	87.1	50.1	94.0
IterDet[9]	88.1	42.2	95.8
CrowdDet[1]	90.3	49.4	-
OneNet[12]	22.7	-	-
DeFCN-POTO[14]	88.7	52.0	96.6
Ours	89.3	50.3	97.3

3.2 Ablation Study

Selection of Sampling Threshold. Because we modeled the distribution of features with Gaussian distribution, we can judge the intensity of features according to the Gaussian response value. In Table ??, the region with Gaussian response value $v > \delta_h$ has strong characteristics, which are hard samples. The areas with response value v below δ_l are mainly the background areas because they are outside the box and have low response value. According to our experiment, we also tried to select segmented samples. The area with δ_h of 0.6 and the area with $v < \delta_l$ are selected. In this way, our samples include not only a certain number of hard samples but also a certain number of easy samples (background). When $\delta_l = 0$, the segmented sampling degenerates into the whole sampling.

We find it necessary to provide a certain number of easy samples for the model because only learning to distinguish the hard samples will lead to unexpected false detection in the easily divided samples. For the segmented sampling method, although its performance is lower than that of the whole sampling method. But we found that it can achieve the highest accuracy in earlier iterations.

Selection of Gaussian radius. Different Gaussian radius can construct different distributions. Therefore, we explored which radius is the best. As shown in Table ??, the radius of Gaussian distribution should be as small as possible, but not less than half of the length and width. Because the distribution value is concentrated on the target due to a small Gaussian radius, so it is unable to sample some parts of the background. And for a large radius, it pays too much attention to the background. In this case, we can solve this problem by changing the sampling threshold.

PDLoss For our PDLoss, we evaluated its speed relative to the Hungarian match. PDLoss can save up to a tenth of time during training. At the same time, positions

Table 2: The influence of different sampling range thresholds in Gaussian balanced sampling. Where $1/r$ is the reciprocal of radius and δ_b is the threshold for selecting the background region ($< \delta_l$), δ_h is the threshold for selecting the foreground region ($> \delta_f$). Where “-” means 0. All the results are relative to the difference between our optimal results in the last line.

Norm	$1/r$	δ_l	δ_h	AP	mMR	Recall
w/o	4	-	0.9	-3.4	+6.3	-1.8
w/o	4	-	0.64	-2.2	+3.8	+0.1
w/o	4	-	0.62	-1.8	+3.0	-0.2
w/o	2	-	1.1	-3.9	+5.9	-1.5
w/o	2	-	0.7	-0.1	+0.5	-1.3
w/o	2	0.2	1.4	-1.0	+1.4	-0
w/o	2	0.1	1.4	-2.5	+4.2	-2.3
w/o	2	0.2	1.2	-0.6	+0.2	-0.6
w	2	-	0.4	-3.6	+6.4	-0.3
w	2	-	0.2	-1.7	+1.7	-0.1
w	2	-	0.1	89.3	50.3	97.3

of the matched sample points are compared. As shown in Figure 2, the simple maximum matching will dominate the loss of targets. Moreover, compared with Hungarian matching, PD loss makes the selected sample points far away from each other, which helps the model better distinguish different targets. But we also need to limit it, it is easy to make the matching sample points too biased because the feature has a large offset. Predicting localization on biased features is no doubt unfriendly for the model. So we add distance clips to limit the effect of loss, and the results are shown in table???. Within a certain range, the loss function works and pushes them as far away as possible.

Table 3: The influence of Non-target Response Suppression. In this experiment, we set $\delta_l = 0$ and $\delta_h = 0.1$. Where “w” and “w/o” means using with and without. All the results are relative to the difference between our optimal results.

Dist-Clip	μ	AP	mMR	Recall
w/o	-	-3.9	+6.2	-0.2
w	10	-1.7	+1.7	+0.2
w	20	89.3	50.3	97.3

As shown in Figure2, (a) is the directly maximum matching method without Hungarian matching or our PDLoss. Compared with other methods, it has more missed detection, although there is no duplicated matching in our example. In fact, it does not distinguish overlapping targets sufficiently because of repetition. That’s why it missed some targets. (b) is the method adopted Hungarian matching. The sample points matched by some targets may be close to each other. This makes the features of the

matching points closer, resulting in insufficient discrimination. (c) is our PDLoss method which can make the matching points far away from each other than before, and the distance loss needs to be limited since the points are too biased shown in (d).

Non-target Response Suppression For the suppression of non-target areas, our main purpose is using it to eliminate false detection. However, to our surprise, it played a crucial role instead. When the non-target area suppression is removed, the AP is greatly reduced as shown in Table 4. This is because the background is easier to distinguish than the target. Weighted loss makes the network pay more attention to positive samples and hard samples. This is also consistent with the consensus that more attention should be paid to hard samples in the traditional one to many methods[11]. In addition, it should be noted that the distribution values we use to weight need to be normalized, otherwise gradient explosion will appear soon.

Table 4: The influence of Non-target Response Suppression(NRS). In this experiment, we set $\delta_l = 0$ and $\delta_h = 0.1$. Where μ_l is the clipped low thresh and non-Norm means do not normalize Gaussian distribution by density. “w” and “w/o” means using NRS or not. All the results are relative to the difference between our optimal results.

NRS	Method		AP	mMR	Recall
	μ_l				
w	Softmax		-1.4	+0.5	-1.3
w	non-Norm		-	-	-
w/o		-	-15.2	+18.0	-0.9
w	clip	0.5	-3.1	+4	-0.5
w		0.1	89.3	50.3	97.3

Since our sampling range δ_h is set to 0.1. Therefore, that we set μ_l to 0 means no lower limit for weight in essence. In addition, in order to prevent the gradient explosion, we also try to use Softmax to normalize the weight. After normalization, the original high response area will still have a higher value than the low response area. The only difference between them is that the loss gets smaller after doing softmax. Finally, if we do not normalize the Gaussian distribution by density and directly multiply it on the loss, the model will face a great risk of gradient explosion.

4 Conclusion

This paper proposes a balanced sampling method based on Gaussian prior distribution and the Non-target Response Suppression method, which effectively solves the problem of positive and negative sample imbalance, and makes the model pay more attention to hard samples. At the same time, PDLoss can effectively avoid the overlapping and dropping of target samples. Our end-to-end framework achieves excellent performance for many of the most advanced NMS detectors and existing NMS-free detectors on the CrowdHuman dataset.

References

1. Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12214–12223, 2020.
2. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
3. Zheng Ge, Zequn Jie, Xin Huang, Rong Xu, and Osamu Yoshie. Ps-rnn: Detecting secondary human instances in a crowd via primary object suppression. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.
4. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
5. Roy Jonker and Ton Volgenant. Improving the hungarian assignment algorithm. *Operations Research Letters*, 5(4):171–175, 1986.
6. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.
7. Songtao Liu, Di Huang, and Yunhong Wang. Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6459–6468, 2019.
8. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.
9. Danila Rukhovich, Konstantin Sofiiuk, Danil Galeev, Olga Barinova, and Anton Konushin. Iterdet: Iterative scheme for objectdetection in crowded environments. *arXiv preprint arXiv:2005.05708*, 2020.
10. Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.
11. Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.
12. Peize Sun, Yi Jiang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Onenet: Towards end-to-end one-stage object detection. *arXiv preprint arXiv:2012.05780*, 2020.
13. Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection, 2019.
14. Jianfeng Wang, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. End-to-end object detection with fully convolutional network. *arXiv preprint arXiv:2012.03544*, 2020.
15. Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018.
16. Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
17. 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.