IQDet: Instance-wise Quality Distribution Sampling for Object Detection

Yuchen Ma, Songtao Liu, Zeming Li, Jian Sun

Megvii Technology

{mayuchen, liusongtao, lizeming, sunjian}@megvii.com

Abstract

We propose a dense object detector with an instance-wise sampling strategy, named IQDet. Instead of using human prior sampling strategies, we first extract the regional feature of each ground-truth to estimate the instance-wise quality distribution. According to a mixture model in spatial dimensions, the distribution is more noise-robust and adapted to the semantic pattern of each instance. Based on the distribution, we propose a quality sampling strategy, which automatically selects training samples in a probabilistic manner and trains with more high-quality samples. Extensive experiments on MS COCO show that our method steadily improves baseline by nearly 2.4 AP without bells and whistles. Moreover, our best model achieves 51.6 AP, outperforming all existing state-of-the-art one-stage detectors and it is completely cost-free in inference time.

1. Introduction

Recent object detection methods have proposed various sampling strategies, assigning the predicted boxes to the ground truth according to their quality. These sampling and assignment strategies have led to great progress in modern object detection task.

As shown in Fig. 1, the improvements in sampling strategies can be divided into two tendencies. (1) From *Static* to *Dynamic*: *Static rules* like RetinaNet [8] assign anchor boxes to the ground truth according to the predefined quality of anchors (*i.e.*, IoU), while *Dynamic rules* (*e.g.*, loss-based sampling in PAA [6]) further boost the performance of the dense object detection, because they are prediction-aware and directly determined by the prediction quality. (2) From *Sample-wise* to *Instance-wise*: Assigned by the *Sample-wise* rule like RetinaNet[8], a predefined anchor is assigned as a positive sample when its IoU is higher than the threshold. Meanwhile, based on *Instance-wise* distribution, [14, 15] propose a center-based distribution according to the ground-truth(GT)'s location, where the samples located inside the central region are assigned as positive samples. These two tendencies achieve better performance and
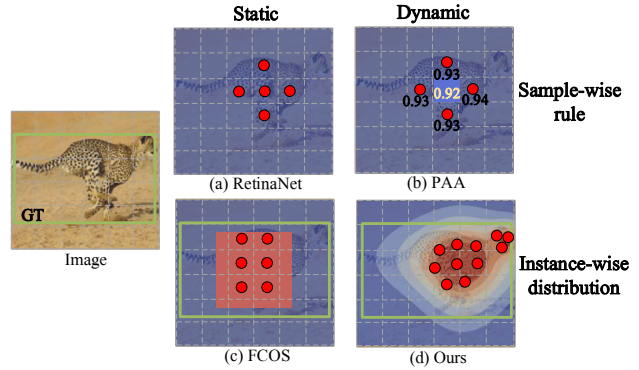

Figure 1. Different sampling strategy of dense object detection. The green squares are GT and the red circles represent the positive samples. (a) *Static* sampling strategies mainly base on the anchors' locations, while the *Dynamic* sampling strategies are prediction-aware. (b) The IoUs of the predicted boxes are quite similar, but the predicted box with IoU=0.92 is compulsively assigned as a negative sample due to the slightly lower IoU, bringing inconsistent noisy samples for the training process. (c) FCOS uses a static center distribution for sampling. (d) IQDet samples training set according to an instance-wise dynamic distribution.

gradually become the de-facto standard in object detection.

However, these sampling strategies might have a few limitations: (1) *Static rules* are not learnable and prediction-aware (e.g. center region and anchor-based), which may be not always the best choice for some eccentric object. (2) Some *Dynamic rules* like PAA might suffer from the noisy samples and per-sample quality rules, without jointly formulating a quality distribution in spatial dimensions, as shown in Fig. 1(b). (3) They sample uniformly over regular grids of image owing to the dense prediction paradigm, which is difficult to assemble enough high-quality and diverse samples. These methods might either *Static* or *Sample-wise*, failing to completely solve these limitations.

In this paper, we propose an instance-wise(GT-wise) quality distribution for sampling to address these issues: instead of assigning each sample independently, the instance-wise sampling strategy selects training samples based on the quality distribution. To achieve dynamic sampling, we

jointly model an instance-wise distribution from the network prediction, making it learnable and prediction-aware. Specially, we optimize the quality distribution encoding function via a variational encoder, approximating the semantic pattern and appearance of the instance. This dynamic quality distribution filters out noisy samples(shown in Fig. 1(b)) and is easier to be learned than previous sampling strategies. Owing to the previous per-pixel assignment depends on IoUs of each predicted boxes, the assignment of PAA [6] is unstable. Our assignment strategy is more robust and extracts the overall instance-wise feature to generate the assignment. And it can sample more high-quality samples, especially the location in the center of the Gaussian.

Further, we use this distribution to guide sample selection in both spatial and scale dimensions, as shown in Fig. 1(d). Proper samples are resampled with higher probabilities and assigned by higher confidences. Meanwhile, different from the traditional per-grid sampling strategy, we sample more positives from high-quality and diverse predictions with a floating-number coordinate, which may avoid the overfitting problem and improve the performance efficiency. The entire process of instance-wise quality assignment and sampling are learnable and thus can be easily optimized by back-propagation.

The contributions of this work are summarized as:

1 Our main contribution is to propose an instance-wise quality distribution, which is extracted from the regional feature of the ground-truth to approximate each prediction's quality. It guides noise-robustly sampling and it is a prediction-aware strategy.

2 Besides, we formulate an assignment and resampling strategy according to the distribution. It is adapted to the semantic pattern and scale of each instance and simultaneously training with sufficient and high-quality samples.

3 We achieve state-of-the-art results on COCO dataset without bells and whistles. Our method leads to 2.8 AP improvements from 38.7 AP to 41.1 AP on single-stage method FCOS. ResNext-101-DCN based IQDet yields 51.6 AP, achieving state-of-the-art performance without introducing any additional overhead.

2. Related Work

Label Assignment in Object Detection. The task of selecting which anchors are to be assigned as positive or negative samples has recently been recognized as a crucial factor that greatly affects the detector's performance [8, 14, 6, 9, 13]. CenterNet [21] and FoveaBox [7], they both use center-sampling strategy to select the positive samples. GuidedAnchoring [15] leverages semantic features to guide the anchor settings and dynamically mitigate the feature inconsistency with a feature adaption module. FreeAnchor [20] uses a bag of *topK* anchor candidates based on IoU for every object and determines positive anchors based on the estimated likelihood. ATSS [19] suggests an adaptive training sample assignment mechanism by the dynamic IoU threshold according to the statistical characteristics of instances. PAA [6, 3] adaptively separates anchors into positive and negative samples according to the detector's learning status in a probabilistic manner, and they propose to predict the Intersection-over-Unions of detected boxes as a measure of localization quality. AutoAssign [22] uses a fully differentiable label assignment strategy to automatically determines positive/negative samples by generating positive and negative weight maps. MDOD [18] proposed a network by estimating the probability density of bounding boxes in an input image using a mixture model and they got rid of the cumbersome processes of matching between ground truth boxes. POTO [16] uses prediction-aware label assignment to enable end-to-end detection. These methods are either *static* or *sample-wise*, which might have a few limitations. We propose a quality distribution for instance-wise sampling strategy, which aims at solving these limitations and boosting the performance of dense object detection.

Sampling Strategies in Object Detection. The most widely adopted sampling scheme in object detection is the random sampling from all candidates. OHEM [12] mines hard examples which have larger losses at each iteration on the fly. Cascade R-CNN [1] proposes that they only sample high-quality prediction to train, the detector is easier to overfitting, due to exponentially vanishing positive samples, which might be harmful to the performance of the detector. And it was shown to avoid the problems of overfitting at training and quality mismatch at inference. SAPD [23] jointly samples the anchor point as a group both within and across feature pyramid levels and soft-weighted anchor points and soft-sampled pyramid levels across all the pyramid levels.

3. Methods

In this section, we first formulate a quality distribution as an instance-wise representation encoding, which is adapted to the semantic pattern of each instance and noise-robust. Next, based on this distribution, we show how to select samples in a probabilistic manner and train with more high-quality samples (Sec. 3.2), which improves the overall performance by a large margin.

3.1. Formulation of Quality Distribution Encoder

As mentioned in 1, previous sampling strategies [14, 15, 19, 6] suffer from these limitations above. Thus, the key idea is to introduce a robust instance-wise distribution,
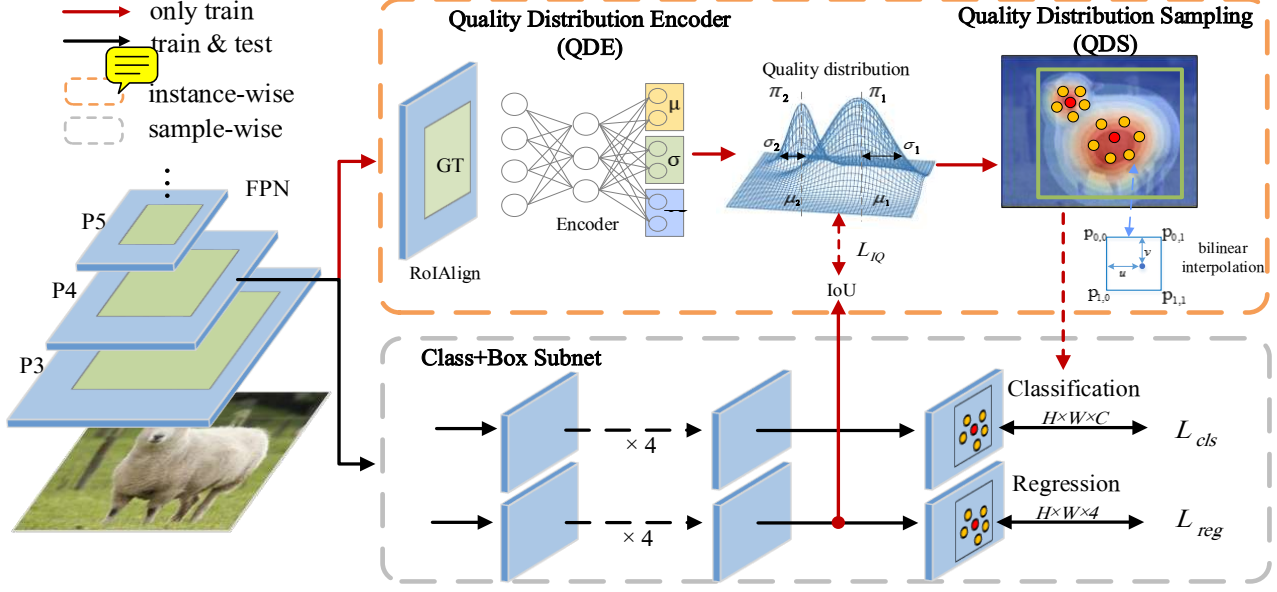
Figure 2. The architecture of IQDet, where the parameters of the quality distribution (μ,σ,π) are predicted by Quality Distribution Encoder (QDE), which represent location, scale, and mixing coefficient of each component respectively. The Class+Box Subnet uses the intermediate output (P3 - P7) from each feature-map of the feature-pyramid, which is similar to FCOS [14]. The subnets in the orange box are the instance-wise operations and the subnets in the gray box are the sample-wise operations. Our sampling strategy is completely cost-free in inference time, as the auxiliary structures only exist during training.

which is prediction-aware and is adapted to the semantic pattern of an object.

We propose a novel distribution learning subnet, called Quality Distribution Encoder(QDE). The architecture of the encoder is demonstrated in Fig. 2. It is worth noting that all operators in Quality Distribution Subnet in the orange box are instance-wise. To effectively encode the instance-wise feature, we first extract the feature of an object according to the GT location and it is realized by applying the RoIAlign layer [4] to each pyramid feature, where the input RoI is the ground-truth box. Specifically, the motivation of using GT feature is that extracting the regional feature of GT is properly aligning with the distribution assignment in spatial dimensions. Thus the encoder is an instance-wise extractor and estimates an overall quality distribution for an instance per pyramid level. To map the GT feature into an instance-wise representation, the encoder should capture the principal component of the ground-truth feature and then estimate an instance-wise quality distribution. Owing to the unknown underlying distribution is not easy to learn, a basic idea is to use an encoder to map the unknown distribution (e.g. quality distribution of per-sample) to a specific distribution like Gaussian.

We choose Gaussian Mixture Model (GMM) to model the distribution. GMM enjoys the several characteristics mentioned in [11]: (1) It can form smooth approximations to arbitrarily shaped distribution. (2) The individual component may model some underlying set of hidden classes.

Thus, according to the GMM modeling, we can estimate a distribution to approximate the sample-wise quality and filter out noisy samples(e.g. the case in Fig. 1(b)). Moreover, each component of GMM may be adapted to a semantic pattern of the instance (e.g. head, body of a person) in spatial dimensions, which may contain more semantic information and high-quality prediction. For each ground-truth $\mathcal{G}_i$, the probability density function of the quality distribution is formulated as follows:

$$p_i(\vec{d}|\mathcal{G}_i, I, \theta) = \sum_{k=1}^{K} \vec{\pi}_{i,k} \Phi_{i,k}(\vec{d}|\mathcal{G}_i, I, \theta), \quad (1)$$

where the parameters K, θ are component number, parameters of the encoder. $\vec{\pi}$ denotes mixing coefficient along x- and y-axis in the spatial dimension of image I, and $\vec{d}$ represents the offsets of the sample inside an object to the GT's center along x- and y-axis. Moveover, we normalize $\vec{d}$ into range [-1, 1] by the ground-truth size. Here, we define each component of the mixture model as: $\Phi_k(\vec{d}) = \mathcal{N}_k(\vec{d}|\vec{\mu}_k, \vec{\sigma}_k) = e^{-\frac{(\vec{d}-\vec{\mu}_k)^2}{2\vec{\sigma}_k^2}}$.

The network architecture of the encoder is shown in Fig. 3(a). For simplicity, the encoder consists of two fully-connected layers, followed by fully-connected layers to predict the learnable parameters. The outputs of learnable encoder are $\vec{\mu}_{n,k}, \vec{\sigma}_{n,k} \in \mathbb{R}^{N \times K \times 2}$ along spatial dimensions and $\vec{\pi}_{n,k} \in \mathbb{R}^{N \times K \times 1}$, which represent the location, scale,
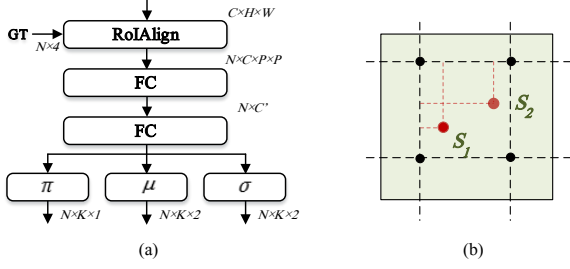
Figure 3. (a) The details of the quality distribution encoder. N, K represent the number of GTs, components of distributions. C, H, W are the feature pyramid sizes. μ, σ denote location, scale of the distribution along x-axis and y-axis, and π is the mixing coefficient. (b) We use bilinear interpolation [5] to compute the exact values of the prediction. Dotted lines are the regular grid of feature map, and green square represent ground-truth $\mathcal{G}$. The red points are the sampling locations of $\mathcal{G}$.

and mixing coefficient of the ground-truth $\mathcal{G}_n$' k-th component respectively. Specifically, the encoder share parameters θ among feature pyramid, and the distributions are estimated separately without feature fusion. To simplify our pipeline, we directly use the latent representation as the quality distribution.

According to the QDE, our goal is to encode the ground-truth feature into quality distribution, which guides the sampling to select more high-quality samples. Thus we supervise the distribution by sample-wise quality metric (e.g. IoU of the predicted boxes). On the one hand, the quality distribution should be learnable and supervised by the quality of each prediction. QDE is a fully differentiable subnet which extracts the GT feature and jointly estimates the overall distribution of an instance. On the other hand, the distribution approximates the sample-wise quality by using GMM, which is easy to learn and noise-robust. Thus, we supervise the quality distribution by the IoU of the predicted boxes over the grid.

3.2. Quality Distribution for Sampling

Based on the quality distribution, we formulate the sampling and label assignment strategy in IQDet. First, instead of sampling over the grid, we resample floating-number coordinate predictions from the distribution. Meanwhile, the distribution represents the quality of these predictions, thus we supervise their scores with the soft distribution.

Resample in Spatial and Scale Dimensions. Previous one-stage detector sample over regular grids of feature map, and they usually select *topK* or IoU thresholds to reject close false positives. However, forcing larger IoU thresholds or smaller *topK* leads to exponentially smaller numbers of positive training samples. They may get involved in the problems of overfitting at training when the ground-truths are assigned with insufficient positive samples.

Based on the distributions mentioned in Sec. 3.1, our sampling strategy select positive samples in a probabilistic manner, and a more diverse and high-quality positive training set is available. First, based on the distribution probability, we resample a fixed number of predictions over floating-number locations, namely sampling continuous coordinates over the feature map. The red circles and orange circles in QDS in Fig. 2 represent the positive samples to be trained. The predictions over these floating-number locations are hard to obtain. Thus, we avoid any quantization of these prediction's locations by using bilinear interpolation [5] to compute the exact values of the predictions. The interpolation operators are calculated over the classification and regression prediction maps. Under the sampling strategy, IQDet is more likely to resample the predictions which are close to the center of the Gaussian. Therefore, we can resample sufficient and high-quality samples instead of sampling over regular grids of feature maps.

According to the bilinear interpolation of the predictions, we construct a bag of candidate predictions for each object $\mathcal{G}$ by resampling K positive predictions in terms of their quality distribution values. Based on the spatial quality sampling above, we use the distribution to select the training positive samples in a probabilistic manner. For the negative samples, our sampling strategy is similar to the baseline. Namely, we select the negative sample outside the GT box over the grid. It should be noted that our sampling strategy is different from the loss reweighting strategy. As shown in 3(b), owing to the classification feature map is a nonlinear distribution, thus the IQDet back-propagates different gradient to the grid point. Moreover, the interpolation for each sample is similar to the RoIAlign [4] when the *pool-size* equals to 1. It is well aligned and preserves the per-pixel spatial correspondence. To sum up, compared with the previous strategies sampling over the grid, IQDet is able to train with more diverse and high-quality samples.

Soft Label Assignment. For each sample selected from the quality distribution, we also calculate the label according to bilinear interpolation. Different from the label assignment in previous work, we use the quality distribution as the labels to supervise the classification prediction. Without dividing the samples into positive and negative samples, IQDet assigns the quality distribution as a soft label to each sample. The predictions with high-quality value are more likely supervised by higher classification target. For predictions located outside the ground-truth region, we assign them as negative samples for sure, which the classification targets are set to zero. It is worth noting that Quality Distribution Sampling (QDS) is applied to all feature pyramid levels. Due to the GMM variables of an instance in each FPN layer are different, the distributions encoded by different pyramid layers are not similar. Namely, for an instance

in different pyramid layers, we assign the label with different quality distributions independently. Across all feature pyramid levels, only the K samples in terms of distribution values are supervised by the distribution, and the others are supervised as zero. For the regression branch, we use the offset of the sampling location with ground-truth as the regression target, which is similar to the previous works. Furthermore, to compensate for the interference caused by the different downsampling rates of FPN, we normalize the distance $\vec{d}$ by its FPN stage's downscale ratio.

3.3. IQDet

Network Architecture. We now present the network architecture of our IQDet. The framework of IQDet is demonstrated in Fig. 2. In our experiments, we adopt a simple anchor-free object detector FCOS [14] as our baseline. The network consists of a backbone, a feature pyramid, and one detection head per pyramid level, in a fully convolutional network. The pyramid levels are denoted as $P_3 - P_7$, which is similar to the baseline.

The Quality Distribution Subnet is shown in the orange box of Fig. 2. First, taking the pyramid feature maps as input, the Quality Distribution Encoder extracts the regional feature of the ground-truth, then it estimates a quality distribution which guides the sampling and label assignment strategy. The operators in this Quality Distribution Subnet are instance-wise, and the samples of an object are estimated and supervised jointly. Moreover, This subnet acts as an auxiliary supervision to the detection backbone during training and it is completely cost-free in inference time. In Class+Box Subnet, the operators are sample-wise, and each sample is estimated and supervised separately. The subnet architecture is exactly the same as the one in FCOS [14] and ATSS [19], which is RetinaNet [8] with modified feature subnets and an auxiliary prediction head. In our experiments, we use predicted IoUs with corresponding GT boxes, instead of centerness scores when using the auxiliary prediction as in [14].

Training and Inference. The proposed IQDet is easy to optimize in an end-to-end way using a multi-task loss. Combining the output of the IQDet, we define our training loss function as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{aux} + \lambda_{IQ}\mathcal{L}_{IQ} \qquad (2)$$

In the implementation, focal loss [8], IoU loss and Binary Cross-Entropy (BCE) loss are used as the classification loss, regression loss and auxiliary loss respectively, which are the same as FCOS [14]. For the $\mathcal{L}_{IQ}$ loss shown in Fig. 2, we supervise the quality distribution by the IoUs of predicted boxes. As the loss function, we use the BCE loss between the target values and the distribution predictions.

Table 1. Ablation study of Quality Distribution Encoder (QDE) and Quality Distribution sampling (QDS). "$*$" denotes centerness as auxiliary task, while others use IoU prediction.

Method	QDE	QDS	AP	AP_{50}	AP_{75}
FCOS$*$			38.7	57.5	41.7
FCOS			39.4	57.9	42.4
IQDet	$\checkmark$		40.4	57.7	44.1
IQDet		$\checkmark$	39.6	57.5	43.0
IQDet	$\checkmark$	$\checkmark$	**41.1**	**58.9**	**44.8**

4. Experiments

4.1. Implementation Details

Following the common practice, our experiments are trained on the large-scale detection benchmark COCO train-val35k set (115K images) and evaluated on COCO val set (5K images). To compare with the state-of-art approaches, we report COCO AP on the test-dev set (20K images).

Training Details. We use ResNet-50 with FPN as our backbone network for all the experiments, if not specified. We use synchronized stochastic gradient descent (SGD) over 8 GPUs with a total of 16 images per minibatch (2 images per GPU) for 90k iterations. With an initial learning rate of 0.01, we decrease it by a factor of 10 after 60k iterations and 80k iterations respectively. Weight decay of 0.0001 and momentum of 0.9 are used. And the input images are resized to ensure their shorter edge being 800 and the longer edge less than 1333 unless noted. In our experiments, λ_{IQ} is 1 and the number of GMM components equals 2, and our method is not sensitive to these values.

Inference Details. During the inference phase, we resize the input image in the same way as in the training phase, and then forward it to obtain the predicted bounding boxes with a predicted class. The following post-processing is exactly the same with RetinaNet and we directly use the same post-processing hyper-parameters (such as the threshold of NMS) of RetinaNet. Finally, the Non-Maximum Suppression (NMS) is applied with the IoU threshold 0.6 per class to yield the final top 100 confident detections per image.

4.2. Ablation Study

4.2.1 Quality Encoder and Quality Sampling.

To demonstrate the effectiveness of the two key components, we gradually add the Quality Distribution Encoder (QDE) and Quality Distribution sampling (QDS) to the baseline to investigate the performance of our proposed IQDet. Based on the FCOS as our baseline, we first analyze the impact of QDE. According to the quality distribution, we just use the soft label assignment and sample over the

Table 2. The ablation experiments of learnable parameters in quality distribution. "✓" means this parameter is learnable.

	$\vec{\mu}$	$\vec{\sigma}$	$\vec{p}$	AP	AP_{50}	AP_{75}
Center Region				38.7	57.5	41.7
IQ-Distribution (GMM)				29.8	47.9	32.1
	✓			38.6	57.2	41.8
	✓	✓		40.0	57.9	43.3
	✓	✓	✓	**40.4**	**58.4**	**43.6**

Table 3. For the second row, we use the RoIAlign to extract feature, then we eliminate the spatial information by a pooling layer. We also compare different pooling size of RoIAlign.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RoIPool [10]	40.6	58.3	44.0	23.5	44.4	53.2
RoIAlign (pool)	40.1	58.3	43.6	23.0	44.0	52.5
RoIAlign [4]	**41.1**	**58.9**	**44.8**	**23.5**	**44.9**	54.4
RoI (3×3)	40.6	58.8	44.4	23.2	44.5	53.7
RoI (7×7)	**41.1**	**58.9**	44.8	23.5	**44.9**	54.4
RoI (14×14)	**41.1**	58.8	**44.9**	23.7	**44.9**	**54.6**

grid of the feature map. As the results in the second row of Table. 1, QDE brings relatively significant performance gain by 1.7 AP, which suggests that the quality distribution is critical for guiding the training. Then we also analyze the effectiveness of QDS in the third row of the table, where the sampling distribution is only based on the center region. We select positive samples with continuous coordinates. The proposed QDS also improves FOCS by 0.9 AP. Finally, the implementation of the IQDet can improve the performance by 2.4 AP in total.

4.2.2 Quality Distribution Encoder.

Learnable Parameters of Quality Distribution. IQDet uses Gaussian Mixture Model (GMM) to estimate the quality distribution, which guides the sampling strategy. To analyze the design of the quality distribution, we compare different quality distributions in Table. 2. To simplify the analysis, we denote the "✓" as learnable parameters and analyze the learnable part of the mixture distribution to verify the gains of each part. We can see that the learnable $\vec{\mu}$ leads to a gain of nearly 10 AP and the learnable $\vec{\sigma}$ can further improve 1.4 AP. Learnable peak $\vec{p}$ of the gaussian distribution shows improvements of 0.4 AP. From the whole table, we can see that the quality distribution brings major improvement and the design of learnable quality distribution is critical.

Architecture of Encoder. To analyze the impact of the RoI feature extractor, we compare the different architecture of feature extractor, shown in Table. 3. We notice that RoiAlign extractor achieves 41.1 AP in the third row, outperforming the counterpart with pooling by 1.0 AP. It demonstrates that the spatial information is crucial for our distribution modeling. We also compare different pooling size of RoIAlign. Considering the speed/accuracy trade-off, pooling size equals 7 in all our experiments.

4.2.3 Quality Distribution Sampling.

Quantified Sampling vs. Continuous Sampling As described in Sec. 3.1, IQDet selects samples according to quality distribution. We conduct several experiments to

Table 4. Analysis of Quality sampling. "†" represents that we select the prediction with bilinear interpolation [5], while the experiments without "†" use quantified locations to sample the prediction to train.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Uniform	40.4	57.7	44.1	22.7	44.1	53.1
QD	40.8	58.6	44.6	23.4	44.4	54.1
QD†(ours)	**41.1**	**58.9**	**44.8**	**23.5**	**44.9**	**54.4**

Table 5. Varying K for the number of positive predictions resampling based on quality distribution.

K	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
4	40.6	58.4	44.0	23.4	44.4	53.8
8	40.6	58.6	44.1	23.1	44.1	53.6
12	**41.1**	**58.9**	**44.8**	23.5	**44.9**	**54.4**
16	40.6	58.3	44.5	23.4	44.5	53.0
20	40.5	58.4	44.0	**23.6**	44.3	53.5

study the effect of the quality sampling strategy. In our experiment, we first select the samples by uniform distribution in spatial dimensions, which is shown in the first row of Table. 4. Then we use quality distribution to sample, but we quantify the locations by rounding off the coordinates. The quality distribution sampling with quantified location show 0.4 AP relative improvement over the baseline. Finally, we use bilinear interpolation to supervise the train samples with continuous coordinates. Specifically, IQDet uses the value of the distribution as the probability to sample, leading to sample more high quality and sufficient samples than other sampling strategies. It increases the performance by 0.7 AP which demonstrating that sampling sufficient and high-quality bounding boxes by QDS will improve detector performance.

Number of Positive Samples of Resampling. Moreover, IQDet selects K positive samples in terms of their quality distribution values. It achieves better performance when the number of positive samples is set to 12, shown in Table. 5.

Quality Distribution Target. We conduct additional ablation studies regarding the target of quality distribution.
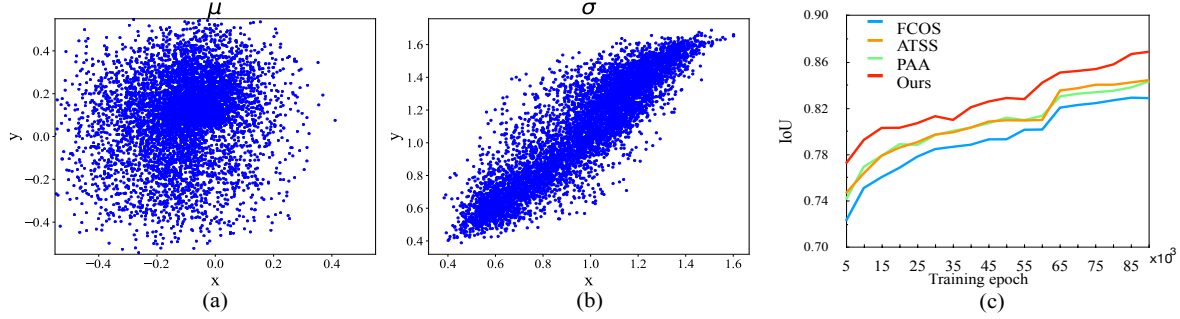
Figure 4. (a)(b) are the μ, σ of Gaussian components respectively and each blue point denotes a quality distribution for an instance. (c) Comparison of IoU curves for different sampling strategies, it suggests that IQDet trains with more high-quality samples.

Table 6. Ablation study of quality distribution target. "Comb-loss" denotes the combined loss of the classification loss and regression loss.

	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Cls-loss	40.7	58.4	44.4	**24.2**	44.1	53.3
Reg-loss	40.6	58.7	43.9	23.1	44.6	54.0
Comb-loss	40.6	58.2	44.2	23.9	44.3	54.0
Ours(IoU)	**41.1**	**58.9**	**44.8**	23.5	**44.9**	**54.4**

The quality of the bounding box can be represented in many forms (e.g. IoU or training loss) to guide the label assign. We compare the performance of detectors with different forms of quality distribution targets. In the first three rows, the predicted boxes with the K smallest L_{cls} and L_{reg} as the positives samples, and we normalize the loss value into the range [-1, 1]. Table. 6 shows that loss-based label assignments are comparable. For the comb-loss, we combine the L_{cls} and L_{reg} to guide the label assign, marked as *comb-loss* in the Table. Finally, the last row of the table is our label assign strategy which uses the IoU of predicted boxes to guide the label assign strategy. We also verify that IoU prediction is more effective than the other forms of the assignments (40.6 AP vs. 41.1 AP).

4.3. Analysis of IQDet

Analysis of Quality Distribution. We evidence our experiments by visualization of distribution representation. We can see that the peaks of the distribution are mainly distributed in some semantic parts of an object, such as the head and trunk of the human in Fig. 5. For the sampling locations, IQDet is able to select more samples with higher quality than the other strategies. We also provide some statistics on the μ, σ of quality distribution. As shown in Fig. 4.2.3(a)(b), each blue point in this figure represents the value of μ or σ for the distribution. The statistics show that sampling over the center region is not always the best choice for an instance in COCO.

Training curves of Quality Distribution Sampling. The train curves of different strategies are depicted in Fig. 4.2.3(c). Obviously, the average IoUs of positive samples' in IQDet are higher than the other methods, and it demonstrates that training with more high-quality samples can boost the performance of the baseline.

4.4. Comparison with State-of-the-art Detectors.

We compare IQDet with other state-of-the-art detectors on MS COCO test-dev set. We adopt 180K training iterations following the previous works. Results are shown in Table. 7. Our IQDet with ResNet-101 backbone achieves 45.1 AP, outperforming all state-of-the-art one-stage detectors with the same backbone as we known. By changing the backbone and training setting the same to other methods, our method consistently surpasses the label assign approaches including ATSS, PAA. By adopting advanced settings, IQDet reaches 51.6 AP, the state of the art among existing one-stage cost-free methods.

4.5. Generalization of IQDet.

To prove the generalization ability of the IQDet, we evaluate our proposed methods with other cost-free one-stage detectors on different detection datasets. For a fair comparison, without modifying any setting of these methods, we only adjust the training setting following the common paradigm of each dataset. As shown in Table. 8, IQDet still has a strong generalization ability on other datasets. Especially for the WiderFace dataset, our method has 1.4 AP gain than all state-of-the-art one-stage detectors on Wider-Face Dataset.

5. Conclusion

In this paper, we proposed an instance-wise quality distribution sampling framework, called IQDet. It is a learnable and prediction-aware sampling strategy according to a mixture model. Our sample selection is guided by a quality distribution which automatically selects samples according to the spatial pattern of the object. This architecture was

Table 7. IQDet vs. the state-of-the-art mothods (single model) on COCO test-dev set. "†" indicates the multi-scale testing

| Method	Iteration	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FCOS [14]	180k	ResNet-101-FPN	41.5	60.7	45.0	24.4	44.8	51.6
FreeAnchor [20]	180k	ResNet-101-FPN	43.1	62.2	46.4	24.5	46.1	54.8
SAPD [23]	180k	ResNet-101-FPN	43.5	63.6	46.5	24.9	46.8	54.6
ATSS [19]	180k	ResNet-101-FPN	43.6	62.1	47.4	26.1	47.0	53.6
PAA [6]	180k	ResNet-101-FPN	44.6	63.3	48.4	26.4	48.5	56.0
IQDet(Ours)	180k	ResNet-101-FPN	**45.1**	**63.4**	**49.3**	**26.7**	**48.5**	**56.6**
FCOS [14]	180k	ResNeXt-64x4d-101	44.7	64.1	48.4	27.6	47.5	55.6
FreeAnchor [20]	180k	ResNeXt-64x4d-101	44.9	64.3	48.5	26.8	48.3	55.9
SAPD [23]	180k	ResNeXt-64x4d-101	45.4	65.6	48.9	27.3	48.7	56.8
ATSS [19]	180k	ResNeXt-64x4d-101	45.6	64.6	49.7	28.5	48.9	55.6
PAA [6]	180k	ResNeXt-64x4d-101	46.3	65.3	50.4	28.6	50.0	57.1
IQDet(Ours)	180k	ResNeXt-64x4d-101	**47.0**	**65.7**	**51.1**	**29.1**	**50.5**	**57.9**
SAPD [23]	180k	ResNeXt-64x4d-101-DCN	45.4	65.6	48.9	27.3	48.7	56.8
ATSS [19]	180k	ResNeXt-64x4d-101-DCN	47.7	66.5	51.9	29.7	50.8	59.4
PAA [6]	180k	ResNeXt-64x4d-101-DCN	48.6	67.5	52.7	29.9	52.2	61.5
IQDet(Ours)	180k	ResNeXt-64x4d-101-DCN	49.0	67.5	53.1	30.0	52.3	62.0
ATSS [19]†	180k	ResNeXt-64x4d-101-DCN	50.7	68.9	56.3	33.2	52.9	62.2
PAA [6]†	180k	ResNeXt-64x4d-101-DCN	51.3	68.8	56.6	34.3	53.5	63.6
IQDet(Ours)†	180k	ResNeXt-64x4d-101-DCN	51.6	68.7	57.0	34.5	53.6	64.5
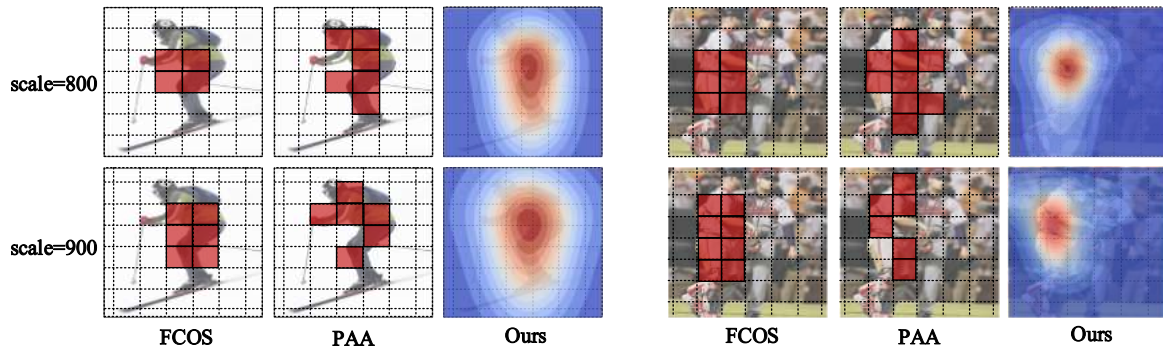
Figure 5. The visualization of label assignment and sampling. The red points are the sampling locations of these strategies.

Table 8. Generalization ability of IQDet. Performance comparison with typical one-stage detectors on PASCAL VOC and WiderFace.

Method	PASCAL VOC [2]			WiderFace [17]		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
FCOS* [14]	56.2	79.9	61.9	50.6	89.1	51.7
ATSS [19]	56.7	79.9	62.3	51.6	89.5	53.9
PAA(w/o vote) [6]	58.3	80.7	64.4	51.0	87.2	53.8
IQDet(Ours)	**59.0**	**81.4**	**65.0**	**52.4**	**89.6**	**55.1**

tasks without additional modification. We believe that it can be useful for many future object detection research efforts.

shown to select more high-quality predictions and avoid the problems of overfitting at training. This method achieves nearly 2.5 AP boost over baseline, and surpassed all previous one-stage object detector on COCO test-dev set. Besides, extensive experiments on other datasets demonstrate that IQDet can conveniently transfer to other datasets and

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[2] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 8

[3] Zheng Ge, Jianfeng Wang, Xin Huang, Songtao Liu, and Osamu Yoshie. Lla: Loss-aware label assignment for dense pedestrian detection. *arXiv preprint arXiv:2101.04307*, 2021. 2

[4] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3, 4, 6

[5] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 4, 6

[6] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. *arXiv preprint arXiv:2007.08103*, 2020. 1, 2, 8

[7] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 2

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007. IEEE, 2017. 1, 2, 5

[9] Han Qiu, Yuchen Ma, Zeming Li, Songtao Liu, and Jian Sun. Borderdet: Border feature for dense object detection. *arXiv preprint arXiv:2007.11056*, 2020. 2

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 6

[11] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741, 2009. 3

[12] 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. 2

[13] Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. Fine-grained dynamic head for object detection. *arXiv preprint arXiv:2012.03519*, 2020. 2

[14] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*, 2019. 1, 2, 3, 5, 8

[15] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[16] 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. 2

[17] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 8

[18] Jaeyoung Yoo, H. Lee, Inseop Chung, Geonseok Seo, and Nojun Kwak. Density-based object detection: Learning bounding boxes without ground truth assignment. *arXiv: Computer Vision and Pattern Recognition*, 2020. 2

[19] 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. 2, 5, 8

[20] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. In *Advances in Neural Information Processing Systems*, pages 147–155, 2019. 2, 8

[21] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2

[22] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*, 2020. 2

[23] Chenchen Zhu, Fangyi Chen, Zhiqiang Shen, and Marios Savvides. Soft anchor-point object detection. *arXiv preprint arXiv:1911.12448*, 2019. 2, 8