Article

LRP-DS: Lightweight RepPoints with Decoupled Sampling Point Set

Jinchao Wang, Libo Weng and Fei Gao *

Laboratory of Graphics & Image Processing, Zhejiang University of Technology, Xihu District, Hangzhou 310000, China; 201426810121@zjut.edu.cn (J.W.); wenglipo@zjut.edu.cn (L.W.)

* Correspondence: gfei_jack@163.com

Abstract: Most object detection methods use rectangular bounding boxes to represent the object, while the representative points network (RepPoints) employs a point set to describe the object. The RepPoints can provide more fine-grained localization and facilitates classification. However, it ignores the difference between localization and classification tasks. Therefore, a lightweight RepPoints with decoupling of the sampling point set (LRP-DS) is proposed in this paper. Firstly, the lightweight MobileNet-V2 and Feature Pyramid Networks (FPN) is employed as the backbone network to realize the lightweight network, rather than the Resnet. Secondly, considering the difference between classification and localization tasks, the sampling points of classification and localization are decoupled, by introducing classification free sampling method. Finally, due to the introduction of the classification free sampling method, the problem of the mismatch between the localization accuracy and the classification confidence is highlighted, so the localization score is employed to describe the localization accuracy independently. The final network structure of this paper achieves 73.3% mean average precision (mAP) on the VOC07 test dataset, which is 1.9% higher than original RepPoints with the same backbone network MobileNetV2 and FPN. Our LRP-DS has a detection speed of 20FPS for the input image of (1000, 600), on RTX2060 GPU, which is nearly twice as fast as the backbone network of ResNet50 and FPN. Experimental results show the effectiveness of our method.

Citation: Wang, J.; Weng, L.; Gao, F. LRP-DS: Lightweight RepPoints with Decoupled Sampling Point Set. *Appl. Sci.* **2021**, *11*, 5876. <https://doi.org/10.3390/app11135876>

Received: 23 May 2021
Accepted: 20 June 2021
Published: 24 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: RepPoints; LRP-DS; decouple; sample; mismatch

1. Introduction

Object detection is one of the most widely used tasks in computer vision. Current object detection methods can be roughly divided into two categories according to whether a priori anchor is needed: anchor-based object detection, and anchor-free object detection. Anchor-based object detection applies a large number of prior anchors to fit the boundary box of a real object, such as the famous single-stage object detection methods of YOLO [1], SSD [2], and the two-stage object detection algorithm represented by Faster R-CNN [3]. The prior anchor is usually set manually according to the statistical characteristics of the dataset. In essence, the anchor can provide a reference for the network, so that the network can obtain some prior knowledge. Therefore, in the training process of network, it is easy to learn the mapping relationship between the anchor and ground truth. In the process of reference, the predicted object box can be obtained according to the prior anchor and the offset calculated by the network. By contrast, anchor-free object detection, which has attracted a lot of attention recently, does not rely on a prior anchor, such as CornerNet [4], ExtremeNet [5], representative points network (RepPoints) [6], etc.

RepPoints directly regards the bounding boxes of object as the regression of the point set. The innovation of RepPoints is to use the set of representation points to describe the object. When the flexible representation method is applied to object detection, the set of representation point can adapt to the geometric changes of the object and provide guidance for feature extraction. However, RepPoints ignores the difference between the object

classification and localization task. Compared with the geometric boundary information of the object, the classification task pays more attention to the semantic key points of the object. For example, for recognition of a cat, the localization task is suitable for extracting features at the boundary of the cat, while the classification task prefers to extract features at semantic key points such as the cat's eyes and nose. In the original RepPoints, the classification and localization tasks share the same sampling positions, which is not appropriate. Therefore, this paper proposes to decouple the feature sampling positions of object classification and localization, so as to give the sampling positions of feature in the classification branch a certain degree of freedom. This allows the feature sampling points of classification to actively find the semantic key position of the object, to improve the recognition accuracy of the classification task.

In the general post-processing method non maximum suppression (NMS), the bounding box with the highest classification confidence is selected in the same category, and then filters out the detection results where the calculated value of the intersection over union (IoU) overlap is greater than the certain threshold. The above steps are repeated. Therefore, in traditional object detection methods, the classification confidence is potentially expected to describe the classification probability and localization accuracy. However, many studies (such as IoU-Net [7]) show that the classification confidence is not enough to describe the localization accuracy of the bounding box. This is considered a mismatch between classification confidence and localization accuracy, due to the double responsibility of classification confidence. On the basis of RepPoints, the problem of mismatch is more serious after the decoupling of feature sampling points between classification and localization. Therefore, the classification confidence is used to describe the category probability of the bounding box, and the localization score is introduced to describe the localization accuracy of the bounding box. In addition, RepPoints uses ResNet50 [8] and Feature Pyramid Networks (FPN) [9] as the backbone network by default, but in order to pursue the lightweight of the model, the lightweight MobileNet-V2 [10] and FPN are used as the backbone network.

Finally, the lightweight RepPoints with the decoupling of sampling point set (LRP-DS) is proposed. The main contributions of this paper are as follows:

- (1) The lightweight MobileNetV2 and FPN are applied as the backbone network to realize the lightweight network;
- (2) The classification free sampling method are proposed in this paper. The sample sets of classification and localization are decoupled, due to the difference between classification and localization tasks;
- (3) The localization score is employed to describe localization accuracy independently, to solve the more serious mismatch of localization accuracy and category probability, after the introduction of the classification free sampling method.

2. Related Work

2.1. Anchor-Based Object Detection

The essence of the anchor-based detection method is to employ a large number of discrete predefined anchors to cover the possible areas of the object, and let the most suitable anchor be responsible for detecting the corresponding object. In 2014, Grishick et al. [11] proposed the first two-stage object detection method, R-CNN. After that, they successively proposed Fast R-CNN [12] and Faster R-CNN [3]. The core of this kind of algorithm is to divide the detection problem into two stages. The candidate regions are generated by the region proposal network (RPN), and then the candidate regions are classified and the localization is modified again. For the mismatch between the classification score and localization result generated in the second stage, Cai et al. [13] proposed Cascade R-CNN network. Cascade R-CNN cascades the R-CNN sub networks of Faster R-CNN many times. In the training, with the cascade level, the IoU threshold of distinguishing positive samples is continuously improved to ensure that the number of samples is not reduced in the case of training high-quality detector. Later, Pang et al. [14] pointed out three imbalances in the training process of Faster R-CNN, and proposed the Libra R-CNN.

Differing from the two-stage method, the single-stage object detection method does not rely on the output candidate box of RPN stage, but directly detects the object. The typical representative of single-stage object detection methods are YOLO [1,15,16], SSD [2] and RetinaNet [17]. SSD is used to detect objects of different scales on six feature maps of different scales. Furthermore, in order to combine the advantages of different scale features, Lin et al. [9] proposed a network FPN using multi-scale pyramid structure to construct a feature pyramid. In addition, Chen et al. [18] think that anchor and feature maps have a mismatch problem in the single-stage detection algorithm, and the two-stage target detection algorithm can alleviate this problem through RoI pooling. Therefore, Chen et al. proposed a special two-stage detection method, AlignDet, by replacing RoI pooling with RoI conv.

2.2. Anchor-Free Object Detection

In the field of object detection, although anchor-based methods are the mainstream, anchor-free detectors show strong vitality due to their efficient performance and anchor-free characteristics. Therefore, anchor-free methods have gradually become a research hotspot. The anchor-free detector can be traced back to YOLO-v1 [1], which outputs the bounding box directly, and does not depend on the prior box anchor. Yu et al. [19] proposed the UnitBox, which uses the current pixel position and the distance between the upper left corner and the lower right corner to form the bounding box of the object. Recent efforts have pushed anchor-free detection to outperform their anchor-based counterparts. Some works, such as CornerNet [4], ExtremeNet [5] and CenterNet [20], reformulate the detection problem as locating several key points of the bounding boxes. To build a bounding box, CornerNet [4] estimates the two corners of bounding box. ExtremeNet [5] represents the bounding box by calculating the four extreme points of the heatmap (top, left, bottom and right) and a center point. CenterNet [20] uses the center point, center point offset, width and height to describe the bounding box. Others like FSAF [21], Guided Anchoring [22] and FCOS [23], encode and decode the bounding boxes as anchor points and point-to-boundary distances. The FSAF [21] network adopts feature selection module (FSAF) based on an anchor-free mechanism to make online feature selection in a feature pyramid. Guided Anchoring [22] can generate anchor automatically in the process of reference. The main contribution of FCOS [23] is to propose a new center-ness branch to reduce the weight of low-quality detection results. In addition, there are some methods exploring the new anchor-free form, such as RepPoints [6]. RepPoints transforms the problem of object detection into a regression problem of the point set, and the bounding box is represented by a flexible representation of the point set. These characterization point sets can describe the precise geometric and semantic features of the object.

2.3. The Mismatch between Classification and Localization Tasks

The source of the mismatch between classification and localization tasks is the difference between the two tasks. The classification task focuses on the semantic information of the object, while the localization task focuses more on the geometric information of the object. The misalignment between classification and localization task can be divided into two aspects. On the one hand, the shared features cause mismatches at the feature level. The shared feature must be able to meet the needs of both tasks, or has a certain compromise. Typically, the popular shared feature of detecting heads in Faster R-CNN [3], brings an increase in speed and a reduction in parameters. However, it forces different tasks to focus on the same feature map. The Double-Head R-CNN [24] proposes to split these two tasks into different heads. It has a fully connected head focusing on classification and a convolution head to pay attention to bounding box regression. The TSD [25] decouples classification and localization from the spatial dimension by generating two disentangled proposals for them. However, it is not appropriate to directly generate two disentangled proposals with weak correlations from the same proposal. It may cause two disentangled proposals to represent different objects.

On the other hand, the classification confidence is not enough to describe the accuracy of localization. During the training process, the classification confidence of the positive samples is expected to be calculated as 1, regardless of whether the IoU between the bounding box and the corresponding ground truth box is high or low. Therefore, the classification confidence does not naturally reflect the localization accuracy. IoU-Net [7] designs an IoU prediction head parallel with the other branch to predict the regressed IoU score. Similarly, MS R-CNN [26] improves Mask R-CNN [27] by attaching a Mask-IoU head parallel with the Mask head to predict the IoU between the predicted mask and the corresponding ground truth mask.

3. Rethinking RepPoints

Usually, a four-dimensional vector $\{x, y, w, h\}$ is applied to represent the bounding box of the object. The (x, y) are the coordinates of the center point of the bounding box, and (w, h) are the width and height of the bounding box respectively. The core idea of RepPoints is to represent the category and location of the object through a more refined point set. Therefore, the RepPoints method describes the object in the form of a point set $P = \{(x_i, y_i) | i = 1, \dots, K\}$. These point sets not only describe the geometric information of the object, but also can be used to guide the feature sampling. Through the loss of localization and classification in the training process, RepPoints can spontaneously find suitable key points for classification and localization. In order to realize this idea, RepPoints employed deformable convolution to sample feature points in point set. Compared with standard regular convolution, the sampling position of deformable convolution can be changed according to the input. Here, the convolution operation can be regarded as the sampling process. Figure 1a is regular sampling grid of standard convolution with a kernel size of 3×3 . In the sampling process of convolution, the positions of nine sampling points present a regular distribution. Figure 1b is a deformable convolution sampling process in which the positions of nine sampling points can be changed dynamically according to the input. In this paper, the sampling center of regular convolution is assumed to be P , and the sampling position is assumed to be R . The R defines the receptive field size and expansion rate of convolution. For example, $R = \{(-1, -1), (-1, 0), \dots, (0, 1), (1, 1)\}$, can represent a sampling position of convolution with kernel size of 3×3 and expansion rate of 1. Then, the sampling set of the regular convolution with the center point at P can be described as $L_r = \{P + R_i | R_i \in R\}$. The deformable convolution employs the offset O to further enhance the sampling position, and its sampling set can be described as $L_d = \{P + R_i + O_i | R_i \in R, O_i \in O\}$.

Figure 1. Illustration of the sampling locations in 3×3 standard and deformable convolutions. (a) Regular sampling grid (green points) of standard convolution. (b) Deformed sampling locations (dark blue points) with augmented offsets (light green arrows) in deformable convolution.

The offset feature map of deformable convolutions can only be learned freely by the network without adding a supervision signal directly. In RepPoints, regression loss is used to directly monitor the generation of the offset feature map, which can make the network learn the desired sampling location explicitly. As shown in Figure 2, the RepPoints can be roughly divided into two steps. In the first step, the object is initially located with the offset

map based on the output feature map of the backbone. The offset map is supervised by regression loss in the training process. Therefore, the offset map represents the point set of predicted objects at each position. The second step is to classify and further fine tune the results of the previous stage, which can use the localization results of the previous stage, the offset map, to guide the sampling of deformable convolution.

Figure 2. Overview of the RepPoints detector.

Most of the current common object detection datasets use regular rectangles to label objects. Therefore, the point set in RepPoints needs to be transformed by pseudo box function F_b , as shown in Equation (1).

$$box' = (x_{min}, y_{min}, x_{max}, y_{max}) = F_b(P), \quad (1)$$

where, box' is a pseudo box; $F_b(.)$ is a pseudo box conversion function, which has three implementation forms: min-max, partial min-max, and moment-based.

In this way, the supervision signal can be directly applied to the pseudo box. In the reference, the conversion function $F_b(.)$ is also used to transform the point set P to the pseudo box to get the final detection result.

In RepPoints, the same set of points are employed to simultaneously describe the feature sampling locations required for classification and localization, which actually ignores the difference between the classification tasks and localization tasks. The localization task is used for bounding box regression, and the geometric information of the object is required for it. The classification task is responsible for the classification of the object, which focuses more on the semantic information of the object than the localization task. Therefore, the same set of sample points is used to meet the needs of the classification and localization tasks, which obviously ignores the differences of the tasks.

In addition, the author of RepPoints proposes three conversion functions F_b : min-max, partial min-max, and moment-based. In fact, the min-max and partial min-max conversion functions use the minimum bounding rectangle of the point set as the bounding box of the object. The essence of the moment-based conversion function is to use the average of all points as the midpoint of the object, and the variance of all points as the width and height of the object. However, there are some problems in the sampling point set with the moment-based conversion function. Figure 3 visualizes the sample points of RepPoints. The green rectangle is the boundary box of the detection object; the starting point of the yellow solid line is the center point, which is the intersection of the yellow solid line in Figure 3, and the ending point of the yellow solid line is the sampling point of the localization branch L_d ; the starting point of the green solid line is the sampling point of

the localization branch L_d , and the ending point is the sampling point after the localization fine adjustment. When using the moment-based conversion function, Figure 3b shows that some sampling points are outside the box of the object, while others are inside the object. This is because the classification task needs to sample the semantic key points of the object, which are generally inside the object, but the localization task needs to calculate the boundary box of the object through the whole point set. Therefore, due to the differences between the two tasks and the characteristics of the moment-based conversion function, some sampling points can only be located outside the object to meet the needs of the two tasks. However, the sampling points outside the target can only collect information irrelevant to the object, which may be unfavorable for object detection. Therefore, the proposed LRP-DS network tends to use the min-max conversion function.

Figure 3. The visualization of sample points for RepPoints, with different convert function: (a) the min-max convert function; (b) the moment-based convert function.

4. The Lightweight RepPoints with Decoupled Sampling Point Set

4.1. Build the Backbone Network Based on MobileNetV2 and FPN

RepPoints adopts Resnet and FPN as the backbone network by default. However, in order to pursue the lightweight of the model, this paper uses MobileNetV2 and FPN as the backbone network. Compared with the backbone network structure of Resnet50 and FPN, the detection accuracy based on MobileNetV2 and FPN is slightly lower, but it has more than double the running speed. The core idea of MobileNetV1 is to use point convolution and depth separable convolution instead of traditional regular convolution. On the basis of MobileNetV1, MobileNetV2 puts forward the inverse residual structure.

The image with resolution 1000×600 is adopted as the input of the network, while the stride of the MobileNetV2 network after removing the full connection is only 32. At this time, the receptive field of the network is relatively limited, which will affect the perception of large object. Therefore, on the basis of MobileNetV2, this paper adds No.8 and No.9 structures in Table 1 to make the network have a greater depth and receptive field. Finally, the adjusted MobileNetV2 structure is shown in Table 1.

Table 1. The struct of adjusted MobileNetV2.

No	Input Channel	Operator	t	c	n	s
1	3	conv2d	-	32	1	2
2	32	bottleneck	1	16	1	1
3	16	bottleneck	6	24	2	2
4	24	bottleneck	6	32	3	2
5	32	bottleneck	6	64	4	2
6	64	bottleneck	6	96	3	1
7	96	bottleneck	6	160	3	2
8	160	bottleneck	6	320	3	1
9	320	bottleneck	6	480	3	2

The conv2d represents a 3×3 regular convolution; the bottleneck represents the inverse residual structure in MobileNetV2; t represents the expansion rate of the number of channels in the inverse residual structure; c represents the number of channels of the output feature; n represents the number of times the structure is repeated; s represents a multiple of reduction relative to the input image.

The core idea of the feature pyramid FPN is to fuse multi-scale information and make multi-scale predictions. It is committed to solving the problem of object scale, and provides a form of top-down information flow. In FPN, the last residual block of conv2, conv3, conv4, conv5 structure in Resnet are defined as $\{C_2, C_3, C_4, C_5\}$ and has $\{4, 8, 16, 32\}$ stride relative to the original image. Then, the FPN fuses the features from top to bottom in order, to get the final feature map $\{P_2, P_3, P_4, P_5\}$. Each feature map P_i is formed by iteratively fusing the contract level feature map C_i and higher level feature graph P_{i+1} . The feature fusion formula is as follows:

$$P_i = W_{3 \times 3}(U_{2 \times}(P_{i+1}) + W_{1 \times 1}(C_i)), \quad (2)$$

where, $W_{3 \times 3}$ and $W_{1 \times 1}$ is a convolution with kernel size of 3×3 and 1×1 , respectively; the function of $W_{1 \times 1}$ is to convert the number of channels into 256; $U_{2 \times}$ is the upsampling operation.

Therefore, adhering to the idea of FPN, the outputs of the last layer of No.7, No.8 and No.9 are regarded as the C_i to be fused and enhanced, and 96 layers are adopted as the channel number of the P_i feature. Thus, the backbone network MobileNetV2 and FPN is formed in this paper.

4.2. Decouple the Sampling Point Set between the Localization and the Classification Tasks

The localization and classification tasks of RepPoints share the same feature sampling location, which ignore the difference of feature concerns between localization and classification. The main idea of this paper is to decouple the task of localization and classification, and give the feature sampling points of classification a certain degree of freedom. The initial point set P obtained by the localization branch is the sampling position set. Due to the supervision signal of the ground truth, it is actually more a geometric boundary characteristic of the object, such as the edge of the object. However, the classification task focuses more on the semantic information of the object, so it is not appropriate for them to share a set of sampling point sets.

Therefore, two classification free sampling methods are proposed which are based on the idea of task decoupling, as shown in Figure 4. The sample set of the original RepPoints can be described as $L_d = \{P + R_i + O_i \mid R_i \in R, O_i \in O\}$. The first scheme, LRP-DS-V1, is shown in Figure 4a. On the basis of the sampling position L_d , a sampling free function for sample point set of classification is introduced to separate the sampling positions of the localization branch and the classification branch. With the influence of classification free sampling, the sampling point set of classification branch is described as $L_c = \{P + F(R_i + O_i) \mid R_i \in R, O_i \in O\}$, where $F(\cdot)$ is the coordinate transformation function based on the degrees of freedom.

Furthermore, this paper designs two kinds of coordinate transformation functions $F(\cdot)$. The two coordinate transformation functions are shown in formula 3 and formula 4.

$$F_1(R_i') = \gamma * R_i' * f_i = \gamma * (x_i' * f_{xi}, y_i' * f_{yi}), \quad (3)$$

$$F_2(R_i) = R_i' + f_i = (x_i' + f_{xi}, y_i' + f_{yi}), \quad (4)$$

where, $R_i' = R_i + O_i = (x_i', y_i')$; $f_i = (f_{xi}, f_{yi})$; f_{xi} and f_{yi} are the free offsets in x -axis and y -axis respectively, and the value range is $[0, 1]$; the γ is the relaxation factor of the coordinate transformation range.

(a) LRP-DS-V1

(b) LRP-DS-V2

Figure 4. Overview of the proposed LRP-DS: (a) the LRP-DS-V1; (b) the LRP-DS-V2.

The set of sampling points on the localization branch is L_d . When the coordinate transformation function $F(\cdot)$ is an identity function, the sampling position L_c based on the degree of freedom degenerates to the original L_d .

Although the design of LRP-DS-V1 can decouple the sampling point set of classification and localization tasks, it may cause the classification task to confuse the object that really needs to be classified, that is, excessive decoupling. Take the left figure of Figure 5 as an example; the yellow rectangle is the current object. Ideally, the sampling position of localization will be distributed at the geometric position of the object. Due to the influence of decoupled freedom, there will be some offset in the sampling position for classification. At this time, the offset of sampling points may lead to the wrong move of classification sampling points to the corresponding object of red rectangle, which will affect the classification of current object.

Figure 5. The examples of perception error for the classification branch of LRP-DS-V1. (a) a perception error between the girl and the horse; (b) a perception error between the cat and the chair.

Therefore, LRP-DS-V2 is proposed in this paper. The basic idea of LRP-DS-V2 is to keep the sampling set L_d to sample the classification branch, but also use the free sampling method to further sample, and finally merge the two sampling results and classify them. It enables the network to perceive the current object that need to be classified, and can also achieve the liberalization of classification sampling points. As shown in Figure 4b, the LRP-DS-V2 firstly uses the sampling set L_d to sample the input feature of the classification branch to obtain the feature A, and then the feature A is performed 1×1 convolution to obtain the free offset feature map of sampling point. Then, the free offset feature map is employed to sample the feature map again on the input feature to obtain feature B. Finally, the feature A and feature B are fused, and the classification is carried out. Feature A indicates the object to be classified, while feature B contains the semantic information currently needed to be classified. Therefore, it can make up for the defects of LRP-DS-V1.

Finally, because the point set is taken as the sampling point of the feature and the idea of task decoupling is proposed in this paper, the pseudo box conversion function of min-max is more suitable for this method than the moment-based pseudo box conversion function.

4.3. Reduce the Mismatch between Classification Confidence and Regression Localization

Non maximum suppression (NMS) is an important part of the current CNN object detection algorithm to remove duplicate bounding boxes. The core of NMS is to select the prediction bounding box with the highest classification confidence in the same category, and eliminate the same category bounding box with IoUs greater than a certain value. Then, NMS performs iteratively in the above manner. This makes the classification confidence naturally assigned two responsibilities of describing the category probability and localization accuracy. However, in the training process, no matter whether the IoU value between the predicted bounding box and the corresponding truth box is high or

low, the classification confidence of positive samples is expected to be one. Therefore, the correlation between the classification confidence and localization accuracy is low, which seriously affects the localization accuracy of the model.

Although the proposed method of decoupling classification and localization tasks can be beneficial to the feature sampling of classification and location, it makes the mismatch between the classification confidence and localization accuracy more serious. Therefore, this paper hopes to introduce a localization score to describe the localization accuracy, so that the classification score only needs to bear the responsibility of category probability. In this paper, for the realization of localization score, the current popular IoU score is selected. The IoU score is obtained by the output of the IoU branch parallel to the location branch, as shown in the location score map in Figure 4. Specifically, the IoU score is obtained by a 3×3 convolution, and the sigmoid activation function is used to normalize the IoU score to $[0, 1]$. In training, this paper applies supervisory signals to IoU scores by only positive samples, but does not use the more robust training methods like IoU-Net. In the process of training, the multi-task loss calculation formula of LRP-DS proposed in this paper is as follows:

$$L = \lambda_1 * L_{cls} + \lambda_2 * L_{init_loc} + \lambda_3 * L_{refine_loc} + \lambda_4 * L_{iou}, \quad (5)$$

where, L_{cls} is the classification loss, L_{init_loc} is the initial localization loss, L_{refine_loc} is the fine-tuned localization loss, and L_{iou} is the IoU score loss.

The classification confidence p describes the category probability of the bounding box, and the IoU score describes the localization accuracy of the bounding box. The object score in the non-maximum suppression NMS can be recalculated by formula (6).

$$Score = p * IoU, \quad (6)$$

5. Experiments

5.1. Experimental Details

The experimental environment of this paper is Intel Core i7 8700 CPU, NVIDIA RTX2070 GPU and Windows10. The experimental comparison is conducted on the public dataset PASCAL VOC [28]. The object detection network is trained on the VOC07 trainval + VOC12 trainval dataset, which contains 16,551 annotated images and 20 predefined object categories. In addition, the object detection network performs performance evaluation on the VOC07 test dataset.

For fair comparison, all experiments are implemented based on Pytorch and mmdetection [29] toolbox. In order to adapt to the more lightweight backbone network MobileNetV2 + FPN, this paper reduces the number of stacked convolutions from four to two, after sampling the classification branches and localization branches of RepPoints and ours LRP-DS. MobileNetV2 uses ImageNet [30] pre-training model for initialization. In mmdetection, the default learning rate is set to 0.01, and the batch size is 16 (eight GPUs and two images are processed on each GPU). Since only one GPU is available and the batch size of each GPU is four images, this paper divides the learning rate by four according to the linear scaling rule to ensure the training effect. The training strategy uses the 1x training method in mmdetection. The training optimizer uses Synchronous Stochastic Gradient Descent (SGD). The image input resolution of the network is scaled by (1000, 600). For data enhancement, only horizontal image flipping is used. Regarding the λ setting of multi-tasking loss, this paper sets $\lambda_1 = 1$, $\lambda_2 = 0.5$, $\lambda_3 = 1$, $\lambda_4 = 0.5$. The relaxation factor γ of the coordinate transformation range is simply set to 1. Compared with RepPoints, the more appropriate minmax conversion function is applied by default to convert point sets to pseudo boxes. In addition, without additional instructions, the backbone network of the object detection network adopts the MobileNetV2 and FPN structure, and all other hyperparameters follow the settings of mmdetection.

5.2. Ablation Study

Table 2 reports the results of ablation study under the PASCAL VOC07 test dataset. From Table 2, it can be found that the method of decoupling the sampling point set proposed in this paper is effective. After applying the degree of freedom decoupling, the mean average precision (mAP) can be increased from 71.4% to 73.1%. Comparing the two coordinate transformation methods proposed in this paper, the coordinate transformation function F_2 is slightly better than the coordinate transformation function F_1 , with a maximum difference of 0.7% mAP. In addition, the experimental results in Table 2 show that LRP-DS-V2 is significantly better than LRP-DS-V1. The second decoupling method proposed in this paper can increase mAP from 72.6% to 73.3%. Therefore, this also verifies that the excessive decoupling of LRP-DS-V1 affects the object perception ability of the classification task, thereby reducing the accuracy. However, the LRP-DS-V2 can make up for this shortcoming.

Table 2. Ablation study on PASCAL VOC07 test set.

Detector	Decouple	IoU	mAP	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat
RepPoints	-	-	71.4	74.0	80.3	73.1	62.7	53.9	79.3	82.9	85.9
	-	✓	71.3	73.0	80.0	73.9	62.1	55.8	79.3	83.0	86.1
	F_1	-	71.7	72.4	79.6	73.8	58.3	56.6	82.4	83.5	85.0
LRP-DS-V1	F_2	-	72.4	75.6	81.5	73.8	63.5	56.3	81.5	83.5	86.2
	F_1	✓	72.0	76.7	78.3	74.2	62.3	56.6	81.0	83.3	84.9
	F_2	✓	72.6	75.9	80.9	74.1	63.3	56.2	82.3	83.3	85.5
	F_1	-	72.5	76.5	79.7	73.9	63.4	57.2	80.4	83.4	85.0
	F_2	-	73.1	77.3	80.9	73.2	64.6	58.1	79.9	83.7	85.6
LRP-DS-V2	F_1	✓	72.8	75.4	79.6	73.7	62.9	56.3	80.8	83.5	85.4
	F_2	✓	73.3	76.4	79.2	74.6	64.3	58.2	81.6	83.4	86.4

In order to verify the necessity of independent description of localization accuracy, the comparative experiment on the IoU branch is conducted. Specifically, this paper designs an IoU branch that outputs localization scores in parallel with the refine positioning branch. The IoU branch only uses a 3×3 convolution, so it hardly affects the amount of calculation of the network. In the training process, in order to simplify, this paper only selects positive samples for training, and does not use more robust training methods. The robust IoU training method is complementary to the method in this paper. Table 2 shows that the RepPoints with IoU branch will not bring any benefits. However, mAP can be further increased to 73.3% for the LRP-DS with IoU branch. This proves the necessity of using classification confidence and localization scores to describe the category probability and localization accuracy of the object, respectively, in our method.

In Table 3, the experimental results of the pseudo box convert function are compared for moment-based and min-max. For the RepPoints, the moment-based pseudo box conversion function has a better mAP. However, for our LRP-DS, the minmax can has a better mAP, with the maximum increase of 1.1% mAP. This is mainly due to the fact that our method puts more emphasis on the point set to guide sampling. Therefore, it is necessary that the point set can better fit the object edge and the semantic key point of the object. The moment-based conversion function has natural shortcomings as shown in Figure 4, so the pseudo box conversion function minmax is more suitable for the LRP-DS object detection network proposed in this paper.

Table 3. The experiment with different convert functions on PASCAL VOC07 test set.

Detector	Convert	mAP	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat
RepPoints	moment	71.9	74.7	78.3	74.6	63.4	53.9	79.3	82.0	85.0
	minmax	71.4	74.0	80.3	73.1	62.7	53.9	79.3	82.9	85.9
LRP-DS-V1	moment	71.7	75.4	79.9	73.8	61.7	53.2	78.7	83.3	85.7
	minmax	72.6	75.9	80.9	74.1	63.3	56.2	82.3	83.3	85.5
LRP-DS-V2	moment	72.2	73.2	79.8	75.6	61.6	54.7	81.4	82.5	86.1
	minmax	73.3	76.4	79.2	74.6	64.3	58.2	81.6	83.4	86.4

Figure 6 visualizes the detection results of LRP-DS-V2. Where, the green rectangle is the boundary box of the detection object; the starting point of the yellow solid line is the center point, and the ending point of the yellow solid line is the sampling point of the localization branch L_d ; the starting point of the green solid line is the sampling point of the localization branch L_d , and the ending point is the sampling point after the localization fine adjustment; the starting point of the red solid line is the sampling point of the localization branch L_d , and the ending point is the sampling point after the degree of freedom conversion of the classification branch.

Figure 6. The visualization of detection results and sample points of LRP-DS-V2.

5.3. Comparison with Other Methods

Tables 4 and 5 show the comparison between our LRP-DS and other methods. In order to make a fair comparison, all object detection networks in the Tables 4 and 5 directly use the implementation code provided by mmdetection. And all networks are retrained using $1 \times$ training strategy in the same environment. The methods of comparison include single-stage object detection method, two-stage object detection method, anchor-free and anchor-based methods. It can be found from Table 4 that, the method proposed in this paper has the best detection performance with the same backbone network of MobileNetV2 and FPN. Table 5 reports the mAP, the GPU memory requirement, multiply-accumulate operation (MACC), the number of parameters, the detection time, and the frames per second (FPS) of the different detectors. The MACC describes the computational complexity of the model. Compared to other detectors, our method has a higher mAP, with similar computational complexity and computational speed. In addition, when backbone of Resnet50 is used, the detection time is 93.2 ms, which is nearly twice as slow as our LRP-DS.

Table 4. The precision comparison with other detectors on PASCAL VOC07 test set.

Detector	mAP	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat
Faster R-CNN [3]	70.3	70.6	79.5	74.0	56.1	46.2	77.5	77.2	85.4
Libra R-CNN [14]	72.5	72.9	81.7	74.4	60.4	51.4	75.4	81.6	85.2
Double Head [24]	72.9	69.6	82.4	71.7	61.1	54.0	80.1	82.9	85.4
Retinanet [17]	70.0	73.4	77.5	72.7	60.3	50.3	76.5	82.3	82.3
FCOS [23]	69.9	74.7	73.9	74.2	63.0	52.1	77.9	82.0	83.4
RepPoints [6]	71.4	74.0	80.3	73.1	62.7	53.9	79.3	82.9	85.9
LRP-DS-V1 (ours)	72.6	75.9	80.9	74.1	63.3	56.2	82.3	83.3	85.5
LRP-DS-V2 (ours)	73.3	76.4	79.2	74.6	64.3	58.2	81.6	83.4	86.4

Due to GPU memory limitation, the batch size of the Double Head is set to 2.

Table 5. The Performance comparison with other detectors on PASCAL VOC07 test set.

Detector	mAP	MACC (G)	Params (M)	GPU Memory (G)	Time (ms)	FPS
Faster R-CNN [3]	70.3	9.16	13.90	1.14	49.55	20.2
Libra R-CNN [14]	72.5	9.18	13.94	1.14	51.28	19.5
Double Head [24]	72.9	26.53	14.69	1.32	73.08	13.7
Retinanet [17]	70.0	8.54	12.34	1.09	45.59	21.9
FCOS [23]	69.9	8.54	12.34	1.09	53.74	18.6
RepPoints [6]	71.4	8.75	12.41	1.14	52.56	19.0
LRP-DS-V1(ours)	72.6	9.02	12.50	1.14	51.08	19.6
LRP-DS-V2(ours)	73.3	9.02	12.50	1.22	52.46	19.1

Due to GPU memory limitation, the batch size of the Double Head is set to 2.

6. Conclusions

This paper proposes a lightweight RepPoints with decoupled sampling point set (LRP-DS). The LRP-DS employs MobileNetV2 and FPN as the backbone network, in order to create a lightweight network and pursue fast detection speed. Considering the differences between classification and localization tasks, two classification free sampling methods, LRP-DS-V1 and LRP-DS-V2, are proposed for decoupling the sampling points of classification and localization. In order to split the responsibility of classification confidence, a localization score is introduced to describe the localization accuracy independently. The final architecture of this paper can achieve 73.3% mAP on the PASCAL VOC07 test dataset, which is better than RepPoints, Libra RCNN and other methods. The experimental results also verify the effectiveness of the proposed method LRP-DS.

Author Contributions: Conceptualization, J.W., L.W. and F.G.; methodology, J.W.; validation, J.W.; formal analysis, L.W.; investigation, L.W.; resources, F.G.; data curation, L.W.; writing—original draft

preparation, J.W.; writing—review and editing, F.G.; visualization, J.W.; supervision, L.W.; funding acquisition, F.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research is being supported by the National Key Research and Development Project of China under Grant No. 2020AAA0104001, the Zhejiang Lab. under Grant No. 2019KD0AD011005, the Zhejiang Provincial Science and Technology Planning Key Project of China under Grant No. 2021C03129 and Scientific Research Fund of Zhejiang Provincial Education Department of China under Grant No. Y201941873.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
2. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
4. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2020**, *128*, 642–656. [[CrossRef](#)]
5. Zhou, X.; Zhuo, J.; Krähenbühl, P. Bottom-Up Object Detection by Grouping Extreme and Center Points. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 850–859.
6. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. RepPoints: Point Set Representation for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 27 October–2 November 2019; pp. 9656–9665.
7. Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; Jiang, Y. Acquisition of Localization Confidence for Accurate Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 816–832.
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
9. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
10. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
11. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
12. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
13. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
14. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 821–830.
15. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
16. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
17. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [[CrossRef](#)] [[PubMed](#)]
18. Chen, Y.; Han, C.; Wang, N.; Zhang, Z. Revisiting Feature Alignment for One-stage Object Detection. *arXiv* **2019**, arXiv:1908.01570.
19. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. UnitBox: An Advanced Object Detection Network. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.
20. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.

21. Zhu, C.; He, Y.; Savvides, M. Feature Selective Anchor-Free Module for Single-Shot Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 840–849.
22. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region Proposal by Guided Anchoring. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2960–2969.
23. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 27 October–2 November 2019; pp. 9626–9635.
24. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking Classification and Localization for Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10183–10192.
25. Song, G.; Liu, Y.; Wang, X. Revisiting the Sibling Head in Object Detector. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11560–11569.
26. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask Scoring R-CNN. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6402–6411.
27. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [[CrossRef](#)] [[PubMed](#)]
28. Everingham, M.; van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [[CrossRef](#)]
29. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
30. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.