TOWARD MINIMAL MISALIGNMENT AT MINIMAL COST IN ONE-STAGE AND ANCHOR-FREE OBJECT DETECTION

Shuaizheng Hao, Hongzhe Liu, Ningwei Wang, Cheng Xu

ABSTRACT

Common object detection models consist of classification and regression branches, due to different task drivers, these two branches have different sensibility to the features from the same scale levels and the same spatial locations. The point-based prediction method, which is based on the assumption that the point with high classification confidence has the high regression quality, leads to the misalignment problem. According to our analyses, the problem is further composed of scale misalignment and spatial misalignment specifically. We aim to resolve the phenomenon at minimal cost—a minor adjustment of the head network and a new label assignment method replacing the rigid one. Our experiments show that, compared to the baseline FCOS, a one-stage and anchor-free object detection model, our model consistently get around 3 AP improvement with different backbones, demonstrating both simplicity and efficiency of our method. Our code will be available at https://github.com/HaoGood/MMMC.

Index Terms—Object detection, feature misalignment, label assignment

1. INTRODUCTION

Object detection is a rather developed research field in the era of deep learning. Two different tasks, classification aiming at researching the differing features across multi classes and regression aiming at drawing the accurate bounding boxes, are often considered. However, due to a huge feature information sensitivity between these two tasks, TSD [1] shows spatial feature misalignment problem exists and compromises the NMS-based models’ ability to predict high-confidence classification and high-quality regression results simultaneously.

Typically, multi-stage detection models [2][3][4] refine a better aligned result by filling the misalignment gap with fully-connected structures in the second stage. Although given a large number of iterations, multi-stage models alleviate the misalignment problem gradually, the cost of a heavy network is far from practical. There is also a difference between anchor-based and anchor-free models in the case of misalignment problem. Compared to anchor-free models, anchor-based models have the choice to assign a single point with one or multiple anchors following IoU assignment rules. During the assignment procedure, anchor-based models have the ability to assign positive labels to the most aligned samples naturally. Thus, we pose a question: is there a simple enough method to solve the misalignment problem, especially in one-stage and anchor-free models?

For this reason, we begin to investigate the fundamental structure of the misalignment problem based on FCOS [5]. Except spatial feature misalignment, we find there is another one, which we call scale feature misalignment, also leading to the phenomenon. To demonstrate the disparity of scale sensitivity in classification and regression branch, we adopt average of classification loss, regression loss and sum of them from each scale level respectively as the scale assignment criteria. The results are in Table 1. When assigned by average classification loss or regression loss alone, models underperform a lot than the model assigned by average sum of them. The disparity of them means there is a myriad of instances whose most acute classification and regression features are not shared by the same scale level. In other words, there is a performance gap between two tasks on the features from the same scale level.

For the spatial misalignment part, we render spatial distributions of classification loss and regression loss within the same instance. As shown in Fig 1, the two distributions are highly misaligned. The points with minuscule classification loss or regression loss have better features for these
Table 1. Models’ performance under different scale assignment standards. “Comb-loss” denotes the combined loss of classification loss and regression loss.

| Standard | AP  | AP_{50} | AP_{75} |
|----------|-----|---------|---------|
| Cls_Loss | 36.7| 56.9    | 39.1    |
| Reg_Loss | 37.2| 55.8    | 40.1    |
| Comb_Loss| 38.6| 57.5    | 41.6    |

two branches to exploit respectively. Thus, the highly misaligned distributions of the loss of two tasks show these two tasks don’t prefer the features at the same spatial locations.

Under these analyses, to solve the scale feature misalignment problem, we design a task-driven dynamic receptive field adaptor (RFA), a simple and plug-in deformable convolution module, as an adjustment of model’s head for each task. To alleviate the negative effect brought by spatial feature misalignment, we design a label assignment method, which is called aligned feature points sampler (ASP), mining the most spatially aligned samples to strengthen the model’s ability to predicate the reliable regression points with high classification scores.

We summarize our contributions as follows:

1. We revisit and decompose the feature misalignment problem in object detection with two components, confirming the existence of scale misalignment problem, in addition to spatial misalignment one.

2. Based on the observation above, we resolve the problems at minimal cost for anchor-free models. Our model, MMC, is in minimal misalignment at minimal cost, and achieves state-of-art performance at 46.3 AP with backbone Res101 evaluated on COCO test-dev set.

2. RELATED WORK

2.1. Scale Misalignment

Since SSD [6] and FPN [7] propose to detect objects with different sizes at different scale levels, there are numerous methods to extract better aligned features for different scale levels. PaNet [8], BiFpn [9] and ASFF [10] merge scale information in a deeper and more complex manner, the performance gap between two tasks is mitigated as each scale level includes more comprehensive information from other scale levels. Recursive-Fpn [11] uses ASPP [12] structure to merge information across scale levels, also mitigating the performance gap between two tasks. Nas-Fpn [13] searches the best formation in a deeper and more complex manner, the performance gap between two tasks is mitigated as each scale level. Recursive-Fpn [11] uses ASPP [12] structure to merge information across scale levels, also mitigating the performance gap between two tasks. Nas-Fpn [13] searches the best formation in a deeper and more complex manner, the performance gap between two tasks is mitigated as each scale level.

3. PROPOSED METHOD

3.1. Dynamic Receptive Field Adaptor

In the heads of modern one-stage detectors, to get feature map with identical size across two branches, each step of four convolutional operations from two branches shares exactly the same kernel size, strid and padding. The receptive field of the last feature map is calculated by:

$$R_f = f(R_i)$$

$$R_{reg} = f(R_{cls})$$

\(R_f\) is the receptive field of initial feature map fed by each FPN level over input image, \(f(\cdot)\) is a static calculation method about receptive field across four sequential convolution layers. Due to the same configurations in the two branches and the static \(f(\cdot), R_{cls}\), and \(R_{reg}\) are identical and static. To drive two branches to take different scale information automatically under the same scale level, as shown in Fig 2 we replace the first convolution procedure with deformable convolution, considering the latter has the ability to dynamically adjust receptive field:

$$R_{cls}^{l} = f(R_i, F_i, \theta_{cls})$$

$$R_{reg}^{l} = f(R_i, F_i, \theta_{reg})$$

at multiple times, thus refining more spatially aligned features for these two different tasks. TSD [1] resolves the problem with lighter cost, by extracting features from two different proposals for classification and regression respectively. However, all these methods fail to apply in modern one-stage detectors.

Regression qualities prediction is also a common-used method. The multiplication result of classification score and regression score is a better standard in NMS procedure compared to classification alone. Iou-Net [14] is the first network learning to predict regression qualities at different proposals. Iou-Aware [15] and PAA [16] employ extra Iou prediction branch to suggest regression-reliable points in one-stage models. Similarly, FCOS [5] has a centerness score prediction branch to represent regression qualities. VFNet [17] merges regression quality prediction branch directly with classification head, by setting Iou scores as soft labels for classification head.

Recently, label assignment is found efficient to improve performance without inferencing cost. ATSS [18] calculates Iou threshold for each instance based on the statistics of Iou between bounding box and anchor boxes. PAA [16] considers the sum of loss from two branches as clustering standard, mining balanced sampling points and anchors for each instance in training procedure. OTA [19] assigns labels in an overall way for each input image, its assignment results are based on optimal transportation theory. These assignment methods inadvertently help model mine spatially aligned points at some degree, but their strategies confine positive samples to instances’ center regions.
problem according to instances’ features dynamically, we design a comprehensive and hyperparameter-single label assignment scheme to mine the most aligned scale levels and spatial locations for each instance sequentially.

**Aligned Scales Assignment Procedure.** For one input image, we define $F_i$ as the feature map fed by each FPN level. For each instance $I_i$ from the input, we define $P_i$ as the feature points within bounding box of $I_i$ from corresponding FPN level. The number of candidate points from each scale level is limited within $K$:

$$N_i = \max(K, \sum_l |P_i^l \in F_i|)$$  \hspace{1cm} (3)

Given candidate numbers $N_i$ for each level, we choose candidate points $C_l$ based on their combined loss of two branches without spatial confinement. The smallest $N_i$ elements are chosen:

$$C_l = \{ \arg \min_{P_i \in P_l, |P_i|=N_i} (L_{cls}^i + L_{reg}^i) \}$$  \hspace{1cm} (4)

As shown in Fig.3, we assign each instance to two scale levels $l^*$ from all levels $L$ to meet the demand of model’s generality on scale dimension. In our case, instead of assigning by IoU thresholds or scale range hyperparameters, we assign directly by the average loss of each scale level forwarded by the same instance $I_i$. The assignment results are chosen by:

$$l^* = \{ \arg \min_{l \in L, ||P_i^l||=2} (\sum_{P_i \in C_l} (L_{cls}^i + L_{reg}^i))/N_i \}$$  \hspace{1cm} (5)

That is given by the intuition: given one instance, the average loss reflects the preference degree for model itself about each scale level. Under this assignment standard, the assigned scale levels are no longer purely driven by instances’ sizes but determined by models’ preference automatically. Further, we take into scale misalignment problem into consideration, as the assignment standard is a balanced sum of classification loss and regression loss. Note that our candidate points $C_l$ are not confined to center regions of $I_i$.

**Aligned Spatial Points Assignment Procedure.** Given scale assignment result $l^*$ for each instance $I_i$, and candidate points $C_l$, from $l^*$, the task for us is to further mine the most spatially aligned points from $C_l$. There are two metrics to be considered for each candidate point: (1) the overall fitness degree $S_f$ under consideration of both tasks; (2) the misaligned degree $S_m$ caused by spatially misaligned loss distributions as we mentioned in section1. As $L_{cls}$ and $L_{reg}$ reflect model’s preference degree for respective tasks naturally, $S_f$ and $S_m$ are evaluated by pair-wise loss ($L_{cls}$, $L_{reg}$) as follows:

$$S_f = \left( \frac{\exp(L_{cls})}{\sum_j \exp(L_{cls}^j)} + \frac{\exp(L_{reg})}{\sum_j \exp(L_{reg}^j)} \right) / 2$$  \hspace{1cm} (6)

$$S_m = \begin{cases} \text{sigmoid}(L_{cls} - L_{reg}) & L_{cls} > L_{reg} \\ \text{sigmoid}(L_{reg} - L_{cls}) & L_{cls} < L_{reg} \end{cases}$$  \hspace{1cm} (7)
Fig. 3. An illustration of aligned feature points sampling procedure. Green dots are candidate points and red dots are positive assignment result. Vectors $L_{cls}$, $L_{reg}$ are pair-wise and chosen from candidate points orderly.

For fitness degree $S_f$, we use softmax function above to evaluate $L_{cls}$ and $L_{reg}$ separately and redistribute them into the same measurable standard, that is given by the advantage softmax function is monotone and the sum of its output is one. For misaligned degree $S_m$, as the difference between $L_{cls}$ and $L_{reg}$ fluctuations a lot between candidate points, we choose sigmoid function to convert it in a rather unified output.

Note that both $S_f$ and $S_m$ should be small for a high-quality positive point. If only assigning labels according to $S_f$, model may be involved in over-fitting problem as only the easiest samples are assigned with positive labels. On the other side, if only assigning labels according to $S_m$, model would fail to converge during training as it would treat candidate points with high but close pair-wise loss as positive. Given the analyses above, we define the final score of candidate points $C_l^*$ as the square root of the multiplication of fitness degree $S_f$ and misalignment degree $S_m$:

$$S = \sqrt{S_f * S_m} \quad (8)$$

The final score $S$ is both balanced and misalignment-aware of two tasks. The decision of positive points $P$ and negative points $N$ from candidate points $C_l^*$ according to $S$ is based on Gaussian Mixture Model (GMM), as shown in Fig[3] Compared to topK manner, GMM has a better performance as it has the ability to dynamically cluster out positive points and negative points according to scores distribution, which is suitable for our appliance.

4. EXPERIMENTS

In this section we conduct our experiments on COCO benchmark. Following the common practice, we use trainval35k set (115k images) as train set and evaluate on coco val set (5k images) for ablation studies. Extensive experiments are evaluated on test-dev set (20k images) to compare with the state-of-art models.

Table 2. Ablation study of Receptive field Adaptor (RFA) and Aligned Points Sampler (APS). "*" denotes centerness as auxiliary branch, while others use Giou prediction.

| method  | RFA | APS | FPS | AP | AP$_{50}$ | AP$_{75}$ |
|---------|-----|-----|-----|----|----------|----------|
| FCOS*   |     |     | 14.7| 38.6| 57.2     | 41.9     |
| FCOS    |     |     | 14.7| 39.4| 57.3     | 42.7     |
| √       |     |     | 14.6| 40.5| 58.0     | 44.3     |
| √ √     |     |     | 14.7| 40.4| 58.2     | 43.9     |
| MMMC    | √   | √   | 14.6| 41.7| 59.3     | 45.1     |

4.1. Implementation Details

We use ResNet-50 pre-trained on ImageNet and FPN as our backbone and neck for our ablation studies, if not specified. Synchronized stochastic gradient (SGD) is employed on 8 GPUs with mini-batch set to 16. The model is trained with 90k iterations. The initial learning rate is 0.01 and is decayed by a factor of 10 after 60k and 80k iterations. Following other common practice, we use Focal Loss [22] and Giou Loss [23] for classification task and regression task respectively. As recent researches [24, 17] shows, centerness has some shortcomings for predicting regression qualities. We simply replace the supervision signal with Giou. As for label assignment scheme, we set $K$ to 9 and GMM’s minimum and maximum score of the candidate points are set by mean of the two components.

4.2. Ablation Study

Effects of individual components. In this section we verify the effectiveness of each component of our proposal. Receptive field Adaptor (RFA) and Aligned Point Sampler (APS) is gradually added to the baseline to make a comprehensive comparison and analyses. As shown in the second row in Table 2, without complex network design and a bunch of extra cost, only RFA brings a significant performance improvement.
### Table 3. Analysis of plug-in position for RFA in head network.

| Position | Cls | Reg | AP  | AP$_{50}$ | AP$_{75}$ |
|----------|-----|-----|-----|-----------|-----------|
| 1        | √   | √   | 39.8| 58.4      | 43.2      |
| 2        | √   | √   | 39.8| **58.3**  | 42.9      |
| 3        | √   | √   | 39.8| 57.8      | 43.0      |
| 4        | √   | √   | **39.9**|58.2  | **43.3**  |

### Table 4. Analysis of loss gap and loss sum between two branches at model’s different iterations. We compare base model with and without RFA.

| iterations | Base_gap | Rfa_gap | reduction | Base_sum | Rfa_sum | reduction |
|------------|----------|---------|-----------|----------|---------|-----------|
| 15k        | 0.169    | **0.167**| 1.2%      | 0.580    | **0.570**| 1.2%      |
| 30k        | 0.156    | **0.153**| 1.9%      | 0.523    | **0.511**| 2.3%      |
| 45k        | 0.141    | **0.139**| 1.4%      | 0.487    | **0.466**| 4.5%      |
| 60k        | 0.150    | **0.146**| 2.7%      | 0.454    | **0.442**| 2.5%      |
| 75k        | 0.137    | **0.134**| 2.2%      | 0.412    | **0.401**| 2.7%      |

Why RFA reduces scale misalignment phenomenon. As we talked about in section 1, there is a performance gap between two tasks on the features from the same scale level. We consider loss gap and loss sum between two tasks from the same scale level reflect models’ adaption ability on scale information. Thus, we collect the average loss gap and loss sum per instance during one training epoch at model’s different iterations. As shown in Table 4, models with RFA module consistently get around 2% relative reduction on the two metrics at different iterations, showing RFA’s ability on conciliating scale misalignment between two tasks without compromising the whole model’s performance.

Scale and Spatial Label Assignment. Our APS is composed of scale assignment and spatial assignment procedures. As Table 5 shows, when these two procedures are separately adopted, the improvement of model’s performance is minor compared to 0.3 AP and 0.5 AP boost respectively. However, when the two procedures are combined, the model reaches 41.7 AP with an absolute 1.2 AP gain. This synergistic effect means each procedure helps the other find the best aligned result on corresponding scale levels and spatial locations respectively.

Visualization of APS’ Spatial Assignment Result. We visualize the spatial distribution of classification and regression loss, and further mark the positive assignment results on the input image in Fig. 4. There are two instances in input image, one is spatially aligned while the other is misaligned. Compared to ATSS [18] and PAA [16], there isn’t significant assignment difference in the spatially aligned instance. While in the spatially misaligned instance, MMMC assigns more aligned points with positive label according to loss distribution of each task, than ATSS [18] and PAA [16] does.

Comparison with State-of-the-art Methods. We compare MMMC with 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 6. Our model, with tiny extra parameters cost, achieves 46.3% accuracy on Res101 backbone, surpassing previous best anchor-based and anchor-free models.

5. CONCLUSIONS

In this paper we revisit and investigate the fundamental structure of feature misalignment problem in object detection. We analyze the problem is composed of scale and spatial one. For scale misalignment, we propose a plug-in and light mod-
Table 6. MMMC vs. the state-of-the-art methods on COCO test-dev set with Res101 backbone. All models are trained in multi-scale scheme.

| Method     | AP  | AP$_{50}$ | AP$_{75}$ | AP$_S$ | AP$_M$ | AP$_L$ |
|------------|-----|-----------|-----------|-------|-------|-------|
| anchor-based|     |           |           |       |       |       |
| F.Anchor$^{[25]}$ | 41.8 | 61.1 | 44.9 | 23.4 | 44.9 | 52.9 |
| SAPD$^{[26]}$ | 43.1 | 62.2 | 46.4 | 24.5 | 46.1 | 54.8 |
| MAL$^{[27]}$ | 43.6 | 61.8 | 47.1 | 25.0 | 46.9 | 55.8 |
| ATSS$^{[18]}$ | 43.6 | 62.1 | 47.4 | 26.1 | 47.0 | 53.6 |
| GFL$^{[23]}$ | 45.0 | 63.7 | 48.9 | 27.2 | 48.8 | 54.5 |
| PAA$^{[16]}$ | 44.8 | 63.3 | 48.7 | 26.5 | 48.8 | 56.3 |
| anchor-free |     |           |           |       |       |       |
| FCOS$^{[5]}$ | 41.5 | 60.7 | 45.0 | 24.4 | 44.8 | 51.6 |
| FSAF$^{[21]}$ | 42.8 | 63.1 | 46.5 | 27.8 | 45.5 | 53.2 |
| IQ-DDET$^{[28]}$ | 45.1 | 63.4 | 49.3 | 26.7 | 48.5 | 56.5 |
| OTA$^{[19]}$ | 45.3 | 63.5 | 49.3 | 26.9 | 48.8 | 56.1 |
| MMMC(ours) | **46.3** | **64.3** | **50.3** | **27.9** | **50.2** | **57.9** |

For spatial misalignment, we propose a novel label assignment scheme mining the most spatially aligned points. Experiments show that our proposal significantly boost model’s performance and the corresponding misalignment problems are alleviated.

6. REFERENCES

[1] Guanglu Song, Yu Liu, and Xiaogang Wang, “Revisiting the sibling head in object detector,” in CVPR, 2020, pp. 11563–11572.
[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” NIPS, vol. 28, pp. 91–99, 2015.
[3] Ross Girshick, “Fast r-cnn,” in ICCV, 2015, pp. 1440–1448.
[4] Zhaowei Cai and Nuno Vasconcelos, “Cascade r-cnn: Delving into high quality object detection,” in CVPR, 2018, pp. 6154–6162.
[5] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, “Fcos: Fully convolutional one-stage object detection,” in ICCV, 2019, pp. 9627–9636.
[6] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, “Ssd: Single shot multibox detector,” in ECCV. Springer, 2016, pp. 21–37.
[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, “Feature pyramid networks for object detection,” in CVPR, 2017, pp. 2117–2125.
[8] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, “Path aggregation network for instance segmentation,” in CVPR, 2018, pp. 8759–8768.
[9] Mingxing Tan, Ruoming Pang, and Quoc V Le, “Efficientdet: Scalable and efficient object detection,” in CVPR, 2020, pp. 10781–10790.
[10] Songtao Liu, Di Huang, and Yunhong Wang, “Learning spatial fusion for single-shot object detection,” arXiv preprint arXiv:1911.09516, 2019.
[11] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille, “Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution,” in CVPR, 2021, pp. 10213–10224.
[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” IEEE TPAMI, vol. 40, no. 4, pp. 834–848, 2017.
[13] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le, “Nas-fpn: Learning scalable feature pyramid architecture for object detection,” in CVPR, 2019, pp. 7036–7045.
[14] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang, “Acquisition of localization confidence for accurate object detection,” in ECCV, 2018, pp. 784–799.
[15] Shengkai Wu, Xiaoping Li, and Xinggang Wang, “iou-aware single-stage object detector for accurate localization,” Image and Vision Computing, vol. 97, pp. 103911, 2020.
[16] Kang Kim and Hee Seok Lee, “Probabilistic anchor assignment with iou prediction for object detection,” in ECCV. Springer, 2020, pp. 355–371.
[17] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf, “Varifocalnet: An iou-aware dense object detector,” in CVPR, 2021, pp. 8514–8523.
[18] 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 CVPR, 2020, pp. 9759–9768.
[19] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun, “Ota: Optimal transport assignment for object detection,” in CVPR, 2021, pp. 303–312.
[20] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin, “Reppoints: Point set representation for object detection,” in ICCV, 2019, pp. 9657–9666.
[21] Chenchen Zhu, Yihui He, and Marios Savvides, “Feature selective anchor-free module for single-shot object detection,” in CVPR, 2019, pp. 840–849.
[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” in ICCV, 2017, pp. 2980–2988.
[23] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in CVPR, 2019, pp. 658–666.
[24] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang, “Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection,” arXiv preprint arXiv:2006.04388, 2020.
[25] Xiaosong Zhang, Fang Wang, Chang Liu, Xiangyang Ji, and Qixiang Ye, “Learning to match anchors for visual object detection,” IEEE TPAMI, 2021.
[26] Chenchen Zhu, Fangyi Chen, Zhiqiang Shen, and Marios Savvides, “Soft anchor-point object detection,” in CVPR. Springer, 2020, pp. 91–107.
[27] Wei Ke, Tianliang Zhang, Zeyi Huang, Qixiang Ye, Jianzhuan Liu, and Dong Huang, “Multiple anchor learning for visual object detection,” in CVPR, 2020, pp. 10206–10215.
[28] Yuchen Ma, Songtao Liu, Zeming Li, and Jian Sun, “Iqdet: Instance-wise quality distribution sampling for object detection,” in CVPR, 2021, pp. 1717–1725.