MutDet: Mutually Optimizing Pre-training for Remote Sensing Object Detection

Ziyue Huang[1], Yongchao Feng[1] Qingjie Liu[1,2*], and Yunhong Wang[1,2]

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[2] Hangzhou Innovation Institute, Beihang University
{ziyuehuang, fengyongchao, qingjie.liu, yhwang}@buaa.edu.cn

1 The pipeline of MutDet

To better understand the detection pre-training and downstream fine-tuning processes, we provide a global view of the pipeline in Algorithm 1.

2 More Experiments

2.1 Rotated Deformable-DETR on DOTA

To further explore the scalability of our MutDet to different detectors, we compare detection pre-training methods on the DOTA-v1.0 based on Rotated Deformable-DETR. The pre-training and fine-tuning settings are consistent with the experiments on ARS-DETR. As shown in Table 1, MutDet achieves the best performance. Compared to the pre-training free, MutDet improves by 1.9 % in AP_{50} and 6.6 % in AP_{75}. Compared to the DETReg baseline, MutDet improves 2.6% in AP_{75}, indicating that MutDet can effectively learn fine-grained discriminative features related to localization.

2.2 Number of Enhancement Layer

We employ a mutual enhancement module combined with multiple layers to mitigate feature discrepancy. As shown in Table 2, increasing the number of layers enhances fine-tuning performance, with three layers achieving optimal results. Adding more enhancement layers facilitates more interaction between object embeddings and the encoder feature, thus aiding in exploring shared knowledge.

2.3 Expand to other detection tasks and models

MutDet can be applied to general object detection and extended to various DETR detectors (we report 3 rotated and 2 horizontal DETRs), effectively improving their fine-tuning performance. Our experiments on mini-coco with the original D-DETR, with 12 epochs for pre-training and fine-tuning, show that

Algorithm 1 The Pipeline of MutDet

Inputs:
Unlabeled dataset D_u for pre-training
Labeled training dataset D_l^{trn} for fine-tuing
Labeled testing/validation dataset D_l^{tst} for evaluation
Pre-trained backbone and Pre-trained SAM
Outputs:
Pre-trained detector, Fine-tuned detector
Process:
1: **Pseudo-labels generation:** Utilize the pre-trained SAM and backbone to generate rotation-aware pseudo-labels for each image in D_u, including boxes, classes and object embeddings. Detail see the Method Section in the main text;
2: **Detection pre-training:** Freeze the detector's backbone and train the other modules (*i.e.*, Mutual Enhancement Module, DETR encoder and decoder, and Auxiliary Siamese Head) in MutDet on D_u with generated pseudo-labels. Finally, we obtain a pre-trained detector;
3: **Fine-tuning:** Initialize a pure detector (without Mutual Enhancement Module and Auxiliary Siamese Head) using the pre-trained detector's parameters, and fine-tune the detector on D_l^{trn}. Finally, we obtain a fine-tuned detector;
4: **Evaluation:** Evaluate the fine-tuned detector on D_l^{tst} and report the performance.

Table 1: Comparison results on DOTA-v1.0 [4]. All adopt Rotated Deformable-DETR [5] as detector and use ResNet-50 as backbone. The results of the test set are reported. '-' indicates pre-training free. **Red**: optimal results. **Blue**: sub-optimal results.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	AP_{50}	AP_{75}
-	70.9	67.9	32.7	55.4	71.9	68.3	77.3	78.7	74.7	81.4	41.2	58.0	54.7	67.2	55.5	63.7	26.4
UP-DETR [2]	80.4	69.2	33.6	59.9	73.7	61.5	76.3	88.0	76.2	81.1	42.5	62.5	54.5	67.5	45.5	64.8	28.6
AlignDet [3]	78.1	64.4	32.3	49.8	73.1	60.9	75.8	88.9	75.6	78.0	40.1	57.7	52.8	65.9	49.4	62.8	27.4
DETReg [1]	84.4	75.7	30.7	54.9	73.5	62.0	74.8	87.2	75.6	79.8	41.0	61.9	59.9	66.8	53.8	**65.5**	**30.4**
MutDet (Ours)	80.3	66.1	35.7	51.1	74.3	68.7	76.5	89.1	76.9	83.0	44.0	55.3	63.2	68.4	51.5	**65.6**	**33.0**

Table 2: Effect of the number of enhancement layer in the mutual enhancement module in MutDet. All models are evaluated on DIOR-R dataset, using DOTA-v1.0 as pre-training dataset, and ARS-DETR as detector.

# Layer	12 Epoch		24 Epoch		36 Epoch	
	AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}
1	64.6	45.2	67.1	47.6	70.2	50.8
2	66.5	47.2	68.3	49.4	70.4	51.2
3 (Default)	**66.9**	**48.1**	**69.8**	**49.9**	**70.7**	**51.2**

MutDet surpasses DetReg by 1.5% in AP_{50}, as demonstrated in Table 3. In **Table 9** of our paper, we test 3 rotated DETR detectors and found out that MutDet achieves the best performance. Besides, we evaluate MutDet on DIOR dataset with original D-DETR and DN-DETR detectors, as shown in Table 4, where MutDet consistently outperforms the baseline.

Table 3: Experiments on mini-coco.

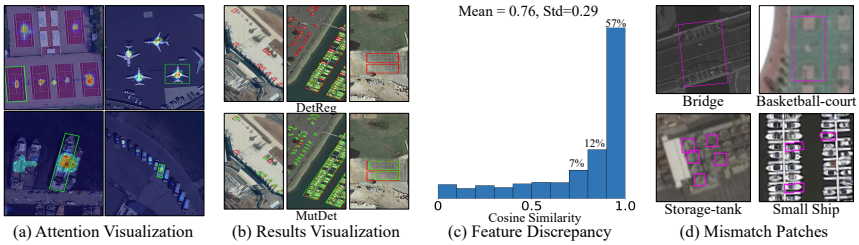
Detectors	Pre-training Methods	mAP	AP_{50}	AP_{75}
D-DETR	-	27.6	43.7	29.2
D-DETR	DetReg	**27.8**	**44.4**	**29.3**
D-DETR	MutDet	**28.1**	**45.9**	**30.1**

Table 4: Experiments with two horizontal DETR on DIOR.

Methods	D-DETR		DN-DETR	
	AP_{50}	AP_{75}	AP_{50}	AP_{75}
-	70.2	49.9	72.4	54.7
DetReg	**73.0**	**54.8**	**74.6**	**57.6**
MutDet	**75.0**	**56.3**	**76.3**	**59.8**

Table 5: Choices of the pseudo-boxes. 'SAM' denotes the SAM-generated boxes. 'GT' denotes the ground truth of dataset. 'SAM & GT' denotes the blend of SAM-genrated boxes and ground truth. Here, we eliminate SAM-generated boxes that have an IoU greater than 0.5 with GT. We pre-train using MutDet on DOTA-v1.0 and then fine-tune on DIOR-R. **Red**: optimal results. **Blue**: sub-optimal results.

Pseudo-boxes	12 Epoch		24 Epoch		36 Epoch	
	AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}
SAM	**66.9**	**48.1**	**69.8**	**49.9**	**70.7**	**51.2**
GT	63.9	45.1	66.2	47.7	69.9	50.5
SAM & GT	**67.2**	**46.8**	**69.1**	**49.3**	**69.9**	**50.7**



(a) Attention Visualization (b) Results Visualization (c) Feature Discrepancy (d) Mismatch Patches

Fig. 1: Qualitative experiments. Zooming in for best view.

3 Further Analysis

3.1 Qualitative experiments

. We have added some qualitative analysis. Figure 1a visualizes the attention weights corresponding to specific query objects (green box), showing that regions with the same class to the query have higher weights, facilitating aggregate semantics. Figure 1b visualizes detection results pre-trained via DetReg and MutDet, leading us to primarily eliminate missing errors. Figure 1c measures the cosine distance between enhanced object embeddings and predicted embeddings

Table 6: Recalls at various IoU thresholds. We compare the class agnostic object proposal performance of three pre-trained models on the DIOR-R trainval set. **Red**: optimal results. **Blue**: sub-optimal results.

Method	IoU Threshold				
	0.5	0.6	0.7	0.8	0.9
SAM	**0.610**	**0.560**	**0.471**	**0.352**	**0.178**
DETReg Pre-trained	0.598	0.522	0.470	0.262	0.088
MutDet Pre-trained	**0.717**	**0.654**	**0.538**	**0.367**	**0.143**

in MutDet, showcasing a notable reduction in feature discrepancy. Furthermore, Figure 1d illustrates the regions with large feature discrepancies, primarily tiny objects or blurred scenes.

3.2 Effect of the pseudo-boxes

Here, We explore the effect of pseudo-boxes on pre-training, Based on MutDet, we vary three strategies to explore pseudo-boxes' impact for pre-training and show the results in Table 5. Although ground truth (GT) boxes are not accessible during the pre-training, we can leverage them for analyzing. Compared to GT, SAM-generated boxes and SAM & GT not only encompass defined classes in the dataset (*e.g.*, vehicles, ships) but also include a substantial number of rare classes (*e.g.* buildings, traffic signs), greatly enriching class diversity. Additionally, removing numerous local pseudo-boxes during the blending process impairs the inter-class diversity of pseudo-boxes. Hence, the detector using SAM-generated boxes as pseudo-boxes achieves the best results.

The diversity of pseudo-boxes is crucial to pre-training performance. This diversity manifests in two aspects: class diversity and inter-class diversity. The former facilitates the model in learning a broader range of visual concepts and features, enhancing its generalization ability. The latter aids the model in capturing variations and subtle features within the same class, thereby improving the model's performance on fine-grained tasks.

3.3 Class Agnostic Object Detection

We evaluate the class agnostic performance [1] of the pre-trained detector. We pre-train for 36 epochs on the DOTA-v1.0 training set and subsequently evaluate the recalls on the DIOR-R trainval set. For pre-trained detectors obtained through MutDet, since object embeddings are not available during testing, we remove the mutual enhancement module and utilize the siamese auxiliary head to obtain the detection results. We apply post-processing to the detection results using NMS with a threshold of 0.7 and a score threshold of 0.1. As shown in Table 6, MutDet achieves higher recall rates than DetReg, surpassing even SAM used for pre-training. Notably, despite DOTA-v1.0 containing only approximately 30,000 images and the backbone being frozen during pre-training,

Fig. 2: Visualization of class-agnostic box predictions on DOTA-v1.0. From top to bottom: (1) Ground truth of the dataset. (2) Predicted boxes by SAM. (3) Predicted boxes by DETReg pre-trained detector. (4) Predicted boxes by MutDet pre-trained detector. DETReg and MutDet are pre-trained on DOTA-v1.0.

MutDet's pre-trained detector still achieves robust class-agnostic detection capabilities and demonstrates strong generalization.

Figure 2 visualizes the boxes predicted by the pre-trained detector. The first and second rows depict the visualization of the ground truth and SAM-predicted boxes, respectively. Compared to DetReg, the pre-trained detector obtained through MutDet exhibits better alignment with SAM detection boxes and fewer cluttered non-object prediction boxes. The comparison indicates that MutDet can effectively learn SAM's class-agnostic detection capabilities and make more efficient distinctions between object and non-object regions.

References

1. Bar, A., Wang, X., Kantorov, V., Reed, C.J., Herzig, R., Chechik, G., Rohrbach, A., Darrell, T., Globerson, A.: Detreg: Unsupervised pretraining with region priors for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14605–14615 (2022)
2. Dai, Z., Cai, B., Lin, Y., Chen, J.: Up-detr: Unsupervised pre-training for object detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1601–1610 (2021)
3. Li, M., Wu, J., Wang, X., Chen, C., Qin, J., Xiao, X., Wang, R., Zheng, M., Pan, X.: Aligndet: Aligning pre-training and fine-tuning in object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6866–6876 (2023)
4. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3974–3983 (2018)
5. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)