

AlignDet: Aligning Pre-training and Fine-tuning in Object Detection

Ming Li^{1*} Jie Wu^{1*†} Xionghui Wang¹ Chen Chen^{2†}
 Jie Qin¹ Xuefeng Xiao¹ Rui Wang¹ Min Zheng¹ Xin Pan¹
¹ByteDance Inc ²Center for Research in Computer Vision, University of Central Florida

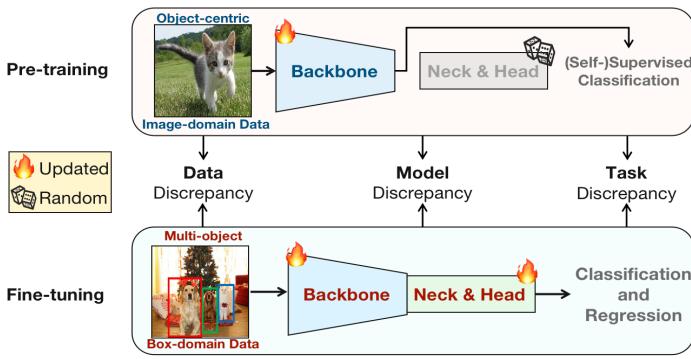


Figure 1. Illustration of the **data**, **model**, and **task** discrepancies between the pre-training and fine-tuning stages in object detection. In mostly pre-training phase, only the backbone is updated via classification task supervision on object-centric datasets such as ImageNet. However, the whole detector is fine-tuned in the multi-objects-based datasets, supervising via classification and regression tasks. By bridging these discrepancies with self-supervised pre-training, AlignDet achieves significant improvements across various **detection algorithms**, **model backbones**, **data settings**, and **training schedules** on COCO. **Project Page:** <https://liming-ai.github.io/AlignDet>.

Abstract

The paradigm of large-scale pre-training followed by downstream fine-tuning has been widely employed in various object detection algorithms. In this paper, we reveal discrepancies in **data**, **model**, and **task** between the pre-training and fine-tuning procedure in existing practices, which implicitly limit the detector’s performance, generalization ability, and convergence speed. To this end, we propose AlignDet, a unified pre-training framework that can be adapted to various existing detectors to alleviate the discrepancies. AlignDet decouples the pre-training process into two stages, i.e., image-domain and box-domain pre-training. The image-domain pre-training optimizes the detection backbone to capture holistic visual abstraction, and box-domain pre-training learns instance-level semantics and task-aware concepts to initialize the parts out of the backbone. By incorporating the self-supervised pre-trained backbones, we can pre-train all modules for various detectors in an unsupervised paradigm. As depicted in Figure 1, extensive experiments demonstrate that AlignDet can achieve significant improvements across diverse protocols,



such as **detection algorithm**, **model backbone**, **data setting**, and **training schedule**. For example, AlignDet improves FCOS by 5.3 mAP, RetinaNet by 2.1 mAP, Faster R-CNN by 3.3 mAP, and DETR by 2.3 mAP under fewer epochs.

1. Introduction

In recent years, there has been significant progress in large-scale pre-training and fine-tuning optimization paradigms in computer vision. A series of pre-training algorithms have been designed to learn domain-sensitive or task-aware concepts to boost downstream performance [19, 24, 1]. As for object detection, current approaches generally leverage ImageNet [14] to pre-train the backbone with classification-oriented supervision. However, compared to the detection-oriented fine-tuning process, this pre-training paradigm exhibits three discrepancies, as shown in Figure 1:

- **Data:** most pre-training methods are conducted on single object-centric datasets, like ImageNet. However, the detection datasets, e.g., COCO [37], usually consist of multiple objects with different scales and locations. The differences in data characteristics and domain can cause the pre-training to deviate from the downstream task.

*Equal contribution. †Corresponding author.

Method	Discrepancy						Generalization			
	Data		Model			Task		Anchor-based	Point-based	Query-based
	Object-centric	Multi-object	Backbone	Neck	Head	Classification	Regression			
Supervised Backbone	✓	✗	✓	✗	✗	✓	✗	✓	✓	✓
MoCo [24], SwAV [6]	✓	✗	✓	✗	✗	✓	✗	✓	✓	✓
DenseCL [57], SelfEMD [38]	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓
PixPro [63], InsLoc [66], SoCo [58]	✓	✓	✓	✓	✗	✓	✗	✓	✓	✗
UP-DETR [12], DETReg [2]	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. We compare with other unsupervised pre-training algorithms in terms of solving the discrepancies and the generalization ability. Taking task discrepancy as an example, previous methods can only introduce the pretext task of classification or regression in the pre-training stage, while AlignDet can build both tasks to learn the semantic and positional information of the objects.

- **Model:** current pre-training algorithms mainly focus on partial modules (such as the backbone) within the model due to the diversity and complexity of detectors. Some key detection components (such as the RPN and regression head) remain random initialization.
- **Task:** existing pre-training approaches only regard the classification task as the pretext task, failing to capture object-aware positional context, including proposal generation, target assignment, and box regression.

These discrepancies potentially bring limited results, poor generalization, and slow convergence speed [48, 52].

As summarized in Table 1, a series of works are proposed to bridge the gap between pre-training and fine-tuning processes in object detection. The initial exploration is to construct dense-level contrastive learning to capture task-sensitive context for dense predictions [57, 38, 63, 66, 58, 59]. Some researchers attempt to pre-train other detection modules, such as FPN [35, 63] and classification head [58]. However, these approaches require hundreds of epochs of pre-training on object-centric datasets and perform poorly when pre-training on COCO. In addition, they neither pre-train all modules, such as the regression head in RetinaNet [36] nor construct appropriate regression pre-training tasks, thus failing to resolve the model and task discrepancies, as illustrated in Figure 2. UP-DETR [12] and DETReg [2] pre-train the entire DETR-like detectors by introducing DETR-sensitive pretext tasks. However, these tasks cannot be applied to other detectors since they rely heavily on the transformer-based decoder. Although the above approaches can alleviate the gap to varying degrees, they cannot comprehensively solve the three discrepancies simultaneously or be generalized to various detectors. This leads us to ask: *how to design a pre-training framework that can address the discrepancies in data, model, and task, and is applicable to all detection algorithms?*

To answer the above question, we propose AlignDet, a universal and pre-training framework for object detection. AlignDet decouples the pre-training process into two stages, image-domain pre-training and box-domain pre-training. The image-domain pre-training optimizes the detection backbone to capture holistic visual abstraction, and the box-domain counterpart learns object-level concepts to

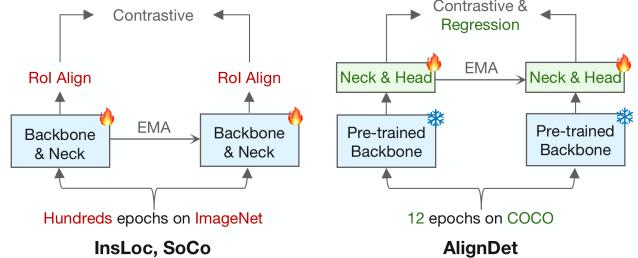


Figure 2. Compared with other box-level pre-training methods (e.g., InsLoc [66] and SoCo [58]), our advantages are: 1) Data: AlignDet works well on COCO with only 12 epochs pre-training; 2) Model: All the modules can be efficiently and fully pre-trained. 3) Task: Both classification and regression knowledge are learned.

initialize the parts out of the backbone. The detector is optimized via box-level contrastive learning and coordinate-related regression losses. It contributes to fully adapting to various detectors, further boosting the performance in the following fine-tuning process, as illustrated in Figure 1. The contributions of this work are summarized in four aspects:

- **New Insight:** We point out that existing detection algorithms are constrained by the data, model, and task discrepancies between pre-training and fine-tuning.
- **Novel Method:** We propose AlignDet to address this issue, which constructs detection-oriented pre-training by learning classification and regression knowledge. It decouples the pre-training into the image domain for the backbone and the box domain for other modules.
- **Efficiency and Pioneering:** The module-based decouple takes full advantage of the existing pre-trained backbones to efficiently pre-train other modules. By incorporating self-supervised pre-trained backbones, we make the first attempt to fully pre-train various detectors using a completely unsupervised paradigm.
- **High Effectiveness:** The comprehensive experiments demonstrate that AlignDet achieves significant performance improvements under various settings, including different detectors, backbones, data settings, and fine-tuning schedules.

2. Related Work

Object Detection. Current object detection algorithms can be divided into anchor-based, point-based, and query-based methods according to different prediction pipelines. The anchor-based approaches generate multiple anchor boxes with pre-defined sizes and scales on each pixel of the feature maps [20, 19, 49, 26, 39, 36]. Usually, they divide the training samples into positives and negatives by Intersection over Union (IoU). The point-based methods aim to find the reference points that correspond to each object, which can be the center points of each instance [72, 53], pre-defined or self-learned key points [31, 73, 67]. Rather than leveraging the pre-defined priors in anchor-based and point-based methods, the query-based methods represent different objects [4, 74, 69] through a set of learnable queries.

Self-supervised Pre-training. Self-supervised learning fully utilizes massive unlabeled data to learn structural data characteristics, and the pre-trained weights are transferred into downstream tasks to ensure good initialization [46, 60, 47, 61, 32, 9, 43, 13]. Numerous pretext tasks have been proposed for unsupervised pre-training, such as feature clustering [5], colorization [30], context prediction [15], rotation prediction [18] and image inpainting [45]. On the one hand, contrastive learning captures good representation by maximizing the similarity of different views from the same instance [24, 10, 8, 22, 11], which achieves competitive performance on multiple downstream tasks. On the other hand, Mask Image Modeling (MIM) has recently attracted increasing attention in self-supervised learning. MIM does not require specific data augmentation and more robust for downstream tasks [1, 71, 64, 23].

Self-supervised Pre-training for Detection. Although unsupervised pre-training has shown competitive results on object detection, there exists a series of inconsistencies in directly transferring image-level pre-trained knowledge to dense-level downstream tasks. To bridge the gap between pre-training and fine-tuning, some approaches propose dense-level contrastive learning to explore the local feature similarity between different views [57, 63, 38, 70, 50]. Some researchers reveal that merely pre-training the backbone is insufficient, and they attempt to pre-training other common modules, such as FPN [35, 58]. However, these methods require expensive pre-training from scratch, and other key modules in detectors (such as the regression head) remain randomly initialized. On the other hand, UPDETR [12] and DETReg [2] pre-train the entire DETR-like detectors by introducing the region matching and feature reconstruction pretext tasks. Although these approaches can pre-train the whole model adequately, the DETR-oriented pretext tasks cannot be directly applied to other detection

paradigms. In contrast, AlignDet achieves efficient and adequate self-supervised pre-training for various detectors.

3. Methodology

Recent works extend the ‘pre-training and fine-tuning’ paradigm via constructing unsupervised pre-training pretext tasks, resulting in higher performance than supervised pre-trained counterparts [24, 63]. However, compared to the detection process, the current pre-training paradigm has inconsistencies in data, model, and task, as illustrated in Figure 1. Although these inconsistencies can be alleviated by training on large-scale labeled datasets [52, 3, 29], it requires enormous computing resources and manual annotation costs. These problems and limitations inspire us to propose AlignDet, a universal and self-supervised framework for bridging the discrepancies between pre-training and fine-tuning in object detection.

The whole pre-training pipeline is summarized in Figure 3. In the following subsections, we introduce the image-domain pre-training and box-domain pre-training in Section 3.1 and Section 3.2, respectively. We provide pseudocode in Algorithm II for a more intuitive understanding of the AlignDet pipeline, and the comparison with other methods on technical details in the **supplementary material**.

3.1. Image-domain Pre-training

Image-domain pre-training optimizes the backbone to extract high-level semantic abstraction for the subsequent box-domain pre-training, as shown in Step 1 on the left of Figure 3. On the one hand, given an image x , the backbone can be pre-trained with a classifier and classification category in the fully-supervised setting. On the other hand, recently arisen unsupervised learning algorithms [24, 10, 11] help to capture more generalized representations with the aid of massive unlabeled data. Take SimSiam [11] as an example, two views x_1 and x_2 are constructed from the input image with different data augmentation. The backbone can learn generalized representations by maximizing the similarity of different views from the same instance. And the predictor and stop gradient are leveraged to prevent the mode collapse [11].

Image-domain pre-training is usually conducted on large-scale image classification datasets such as ImageNet [14], in which each sample mainly contains one or a few salient objects in the center of the image. *It exists an apparent gap because the pre-training procedure has no access to the target domain data, which often contains multiple objects with different scales and locations. Furthermore, the detection head is still randomly initialized and the regression task is not explicitly learned in this image-domain pre-training. To this end, we design the box-domain pre-training to bridge these discrepancies.*

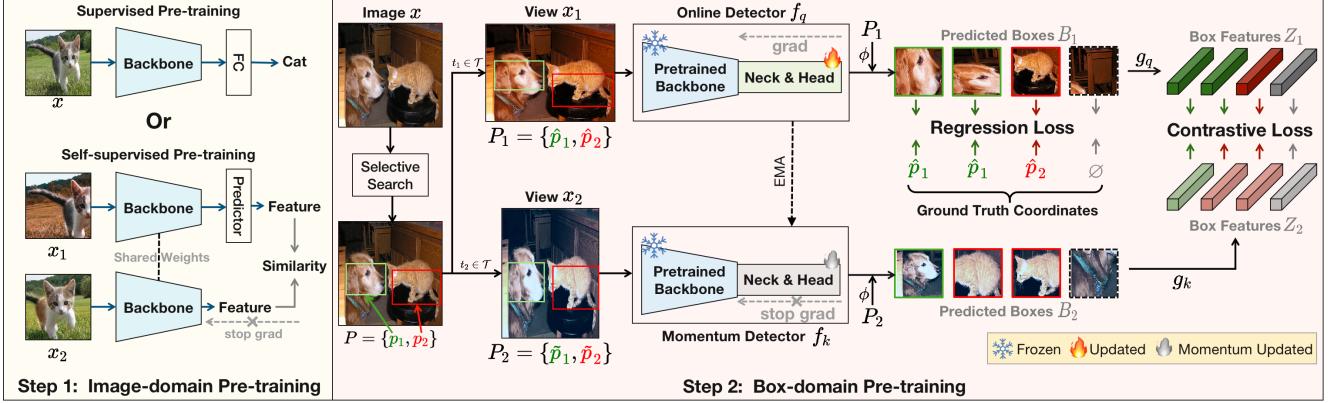


Figure 3. The pre-training pipeline of AlignDet. Both supervised and self-supervised pre-training can be employed in the image-domain stage to capture holistic visual concepts. For the box domain pre-training, selective search is first adopted to generate unsupervised proposals as pseudo labels, then each proposal is augmented to construct two views with different scales and transformations. Each predicted box is used to construct contrastive learning and coordinated-related losses for adapting to detection-oriented tasks.

3.2. Box-domain Pre-training

Self-supervised Object Detection. As represented in Figure 3, given an input image x , we first generate the unsupervised proposals $P = \{p_1, p_2, \dots, p_n\}$ via selective search [54]. Each proposal $p_i \in P$ can be presented as a bounding box with coordinates $(\hat{x}, \hat{y}, \hat{w}, \hat{h})$, where (\hat{x}, \hat{y}) denotes the coordinates of the box center and (\hat{w}, \hat{h}) denotes the width and height. Then the image x is transformed into augmented views x_1 and x_2 by applying a transformation t sampled from the set of image transformations \mathcal{T} . The coordinates of unsupervised proposals P also change into P_1 and P_2 according to the corresponding image transformations. In every iteration of the pre-training procedure, the online detector f_q and momentum detector f_k will generate their predicted boxes of the image at different views. By regarding the unsupervised proposals P_1, P_2 as ground truth, we can obtain the predicted boxes B_1, B_2 by:

$$B_1 = \phi(f_q^{reg}(x_1), P_1), \quad B_2 = \phi(f_k^{reg}(x_2), P_2) \quad (1)$$

where $f^{reg}(x)$ denotes the box coordinates predicted by the regression-related modules f^{reg} based on image x . ϕ represents the target assignment operation, which conducts sample matching between $f_q^{reg}(x_1)$ and unsupervised proposals P_1 , such as calculating IoU in anchor-based detectors. Note that the label-matching method varies with different detectors and we keep the default assignment mechanism of the detector without any changes, ensuring our method can be easily applied to different types of detectors. Different from assigning the category of the matched object in the dataset and calculating classification loss in formal supervised training, we assign $l \in \{\emptyset, 1, \dots, n\}$, the index of

Some anchor-free detectors [53] use pre-defined points to accomplish the procedure and some query-based detectors [4, 74] divide positives and negatives by minimizing regression and classification cost.

each proposal in P to the paired output of online detector f_q and momentum detector f_k by ϕ . Here \emptyset stands for the background, which means that the output box did not match any proposal in P . Then each output box in B_1 and B_2 can be rewritten as $b = (x, y, w, h, l)$.

Box-domain Contrastive Learning. After obtaining the coordinates of each predicted box, we can further obtain the corresponding features in different views by replacing the original supervised classification head in the detection head with the unsupervised head f^{con} :

$$Z_1 = g_q(f_q^{con}(x_1, B_1)), \quad Z_2 = g_k(f_k^{con}(x_2, B_2)) \quad (2)$$

where $f^{con}(x, B)$ denotes the extracted features of predicted boxes B for contrastive learning in the box-domain pre-training procedure. And g is the feature projection module, here we follow the architecture in MoCo v2 [10], *i.e.*, a 2-layer MLP head with ReLU [21]. Please note that f^{con} and f^{reg} are usually two different modules in the detection head, and here we do not draw in Figure 3 for brevity.

In this unsupervised pre-training procedure, contrastive learning is formulated by the principle that the box representation corresponding to the *same* proposal should be similar and vice versa. Specifically, we define the set of query boxes as $Q = \{b \in B_1 : l \neq \emptyset\}$. For each query box $q \in Q$, assuming its assigned proposal index is i and the feature is z_q . The set of positive keys K_+ and negative keys K_- for query feature z_q can be constructed as:

$$Z_+ = \{z \in Z_1 : l = i\}, \quad Z_- = \{z \in Z_1 \cup Z_2 : l \neq i\} \quad (3)$$

Then the box-domain contrastive loss \mathcal{L}_{con} for all query boxes in Q is defined as:

$$\mathcal{L}_{con} = - \sum_q \sum_{z_+} \log \frac{\exp(z_q \cdot z_+ / \tau)}{\exp(z_q \cdot z_+ / \tau) + \sum_{z_-} \exp(z_q \cdot z_- / \tau)} \quad (4)$$

where τ is a temperature hyper-parameter that controls the difficulty of the task of contrastive learning [56], we set 0.5 in our paper for all methods.

Overall Loss. For a prediction box $q_i = (x, y, w, h, l)$, where $q_i \in Q$. l is the index of its corresponding proposal in P_1 which calculated by the label assign function ϕ . And we define it as $\hat{p}_l = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$. The coordinate-related regression losses are calculated according to each detector, and we do not make any modifications during pre-training:

$$\mathcal{L}_{reg} = \sum_q \sum_{k=1}^K \lambda_k \cdot \mathcal{L}_k((x, y, w, h), (\hat{x}, \hat{y}, \hat{w}, \hat{h})) \quad (5)$$

where K is the number of coordinate-related losses, and λ_k denotes the loss factor. \mathcal{L}_k could be any regression loss like IoU loss in FCOS [53] or L1 loss in Mask R-CNN [26]. Then the overall loss is the combination of box-level contrastive loss \mathcal{L}_{con} and the coordinate-related losses \mathcal{L}_{reg} :

$$\mathcal{L} = \lambda_{con} \cdot \mathcal{L}_{con} + \lambda_{reg} \cdot \mathcal{L}_{reg} \quad (6)$$

where $\lambda_{con}, \lambda_{reg}$ are the loss hyper-parameters, and we keep the same with the default setting from the corresponding detector. We freeze the backbone in the box-domain stage to avoid detectors over-fit to noisy pseudo labels. Box-domain pre-training addresses data and task discrepancies by employing multi-object data to construct detection-oriented pretext tasks. By cooperating with image-domain pre-training, AlignDet contributes to pre-train all modules within the detector, thus solving the model discrepancy.

4. Experiments

4.1. Settings

Datasets. In the image-domain pre-training stage, both the ImageNet [14] and COCO [37] dataset can be used to optimize the backbone. And the box domain pre-training employs the box domain datasets that contain non-object-centric and multi-object images. Unless otherwise specified, we pre-train all methods on the COCO train2017 dataset [37] without any labels, then evaluate the detection model on the COCO val2017 dataset. We also follow [10] to fine-tune the detectors on the VOC 07+12 train-val set and evaluate on the VOC 07 test set.

Data Augmentation. For the box-domain pre-training data augmentation, we follow SoftTeacher [65] but remove the RandomCrop and other box-jitter transforms to ensure all objects exist in both views. The image resolution is [640, 800] to construct the multi-scale proposals and objects in the pre-training procedure. In the fine-tuning phase, we follow the default augmentation settings of different methods in mmdetection [7]. And we follow [24, 22, 58, 63, 57, 6] to use multi-scale training for self-supervised backbones.

Algorithm 1 AlignDet Pseudocode, PyTorch-like

```

# x: input images
# p: selective search proposals
# aug: independent random augmentation

for x, p in data_loader:
    (x1, p1), (x2, p2) = aug(x, p), aug(x, p) # augmentation
    x1, x2 = backbone(x1), backbone(x2) # frozen backbone
    x1, x2 = neck_q(x1), neck_k(x2) # feature pyramid

    # proposals as pseudo labels, boxes are predicted
    b1, b2 = head_q.f_reg(x1, p1), head_k.f_reg(x2, p2)
    z1, z2 = head_q.f_con(x1, b1), head_k.f_con(x2, b2)
    z1, z2 = g_q(z1), g_k(z2) # feature projection

    L = loss_con(z1, z2) + loss_reg(b1, b2, p1, p2) # losses
    ema_update(neck_q, neck_k, head_q, head_k)

```

Pre-training and Fine-tuning Details. In the pre-training phase, we fix all parameters of the backbone and update the neck and head. Other hyper-parameters and settings are set as default in mmdetection [7] except the hyper-parameters of prediction sampling. In the fine-tuning phase, all the pre-trained weights except for the projection module are used to initialize the object detection model. Synchronized batch normalization [28] is used in both backbone and FPN following previous work [24, 22, 58, 25], which helps calibrate magnitudes for pre-trained models [24]. All pre-training follows the default 1x (90k) schedule except 50 epochs for DETR. We adjust the learning rate and weight decay following previous work [58]. If not specified, the supervised pre-trained ResNet-50 [27] in PyTorch [44] is used by default for both the pre-training and fine-tuning stages. The hyper-parameter details of different methods and experiments are summarized in the supplementary material.

4.2. Experimental Results

Detectors and Data Settings. To demonstrate the generalization of our approach, we leverage AlignDet to pre-train different detectors, including FCOS [53] (single-stage and point-based), RetinaNet [36] (single-stage and anchor-based), Faster R-CNN and Mask R-CNN [26] (two-stage and anchor-based), and DETR [4] (query-based). Here Faster R-CNN uses ROI Align following previous work [7, 24], and Mask R-CNN is fine-tuned with both detection and instance segmentation annotations following [58]. We random sample 1%, 5%, 10%, and 20% images from COCO train2017 set as the fine-tuning data. We provide 5 different data folds for each low-data setting, and the final performance is the average of all results. To avoid over-fitting and demonstrate the advantage of faster convergence, besides the standard 1x (90k iterations) training schedule, we also report the results of 12k iterations in a low-data regime following [62]. Table 2 summarizes the detection performance of various detectors with different data settings.

Compared with ImageNet pre-training, our AlignDet has significantly improved performance under different settings. Even with complete data (100%) and fine-tuning schedule (90k), there are nearly 1.0 mAP improvements for

Detector	Align	COCO 12k					COCO 90k				
		1% Data	5% Data	10% Data	20% Data	100% Data	1% Data	5% Data	10% Data	20% Data	100% Data
FCOS	✗	8.1	16.5	21.5	23.6	22.1	7.3	16.0	20.0	23.8	36.6
FCOS	✓	9.5 (+1.4)	18.6 (+2.1)	23.3 (+1.8)	26.3 (+2.7)	27.4 (+5.3)	8.5 (+1.2)	17.3 (+1.3)	21.1 (+1.1)	25.1 (+1.3)	37.5 (+0.9)
RetinaNet	✗	8.0	17.8	21.0	23.0	24.2	8.3	18.0	22.2	25.9	36.3
RetinaNet	✓	9.8 (+1.8)	19.3 (+1.5)	23.5 (+2.5)	25.9 (+2.9)	26.3 (+2.1)	9.9 (+1.6)	19.0 (+1.0)	23.1 (+0.9)	26.6 (+0.7)	37.3 (+1.0)
Faster R-CNN	✗	9.2	18.7	24.2	26.6	27.3	6.5	14.0	18.8	24.1	37.9
Faster R-CNN	✓	11.7 (+2.5)	21.2 (+2.5)	26.9 (+2.7)	29.6 (+3.0)	30.6 (+3.3)	8.9 (+2.4)	16.2 (+2.2)	20.1 (+1.3)	25.2 (+1.1)	39.0 (+1.1)
Mask R-CNN	✗	8.8	19.1	24.2	26.5	27.2	7.6	15.2	20.0	25.2	38.3
Mask R-CNN	✓	12.4 (+3.6)	22.4 (+3.3)	27.4 (+3.2)	30.1 (+3.6)	30.5 (+3.3)	9.5 (+1.9)	16.6 (+1.4)	20.8 (+0.8)	25.9 (+0.7)	39.4 (+1.1)
Detector	Align	COCO 50 epochs					COCO 100 epochs				
DETR	✗	7.6	18.9	24.1	29.7	35.0	7.7	19.5	25.0	30.4	38.4
DETR	✓	11.2 (+3.6)	21.4 (+2.5)	26.1 (+2.0)	30.9 (+1.2)	37.3 (+2.3)	10.7 (+3.0)	21.2 (+1.7)	26.0 (+1.0)	31.2 (+0.8)	39.0 (+0.6)

Table 2. With only 12 epochs pre-training on COCO for modules out of backbone, AlignDet achieves consistent improvements across different detectors, training strategies, and data sizes. All the results are conducted with a supervised pre-trained ResNet-50 backbone.

Backbone	Align	Schedule	AP ^{bb}	AP ^{mk}
MobileNet v2	✗	1x	30.1	27.2
MobileNet v2	✓	1x	31.3 (+1.2)	27.9 (+0.7)
ResNet-18	✗	1x	34.5	31.1
ResNet-18	✓	1x	35.7 (+1.2)	31.9 (+0.8)
ResNet-50	✗	1x	38.3	34.3
ResNet-50	✓	1x	39.4 (+1.1)	35.3 (+1.0)
Swin-Small	✗	1x	46.6	41.5
Swin-Small	✓	1x	47.5 (+0.9)	41.8 (+0.3)
Swin-Small	✗	3x	48.2	43.2
Swin-Small	✓	3x	49.1 (+0.9)	43.4 (+0.2)
Swin-Base	✗	1x	48.8	43.1
Swin-Base	✓	1x	49.6 (+0.8)	43.6 (+0.5)
Method	Align	Schedule	AP ^{bb}	AP ^{mk}
SimMIM [†] (Swin-B)	✗	3x	51.0	45.1
SimMIM (Swin-B)	✓	3x	51.6 (+0.6)	45.8 (+0.7)
CBNet v2 [†] (Swin-L)	✗	1x	57.3	49.7
CBNet v2 (Swin-L)	✓	1x	57.8 (+0.5)	50.1 (+0.4)

Table 3. AlignDet achieves consistent improvements on different backbones and the SOTA detector CBNet v2. [†] denotes our reproduced results with official code or models.

Detector	Align	AP	AP ₅₀	AP ₇₅
FCOS	✗	52.6	79.6	57.1
FCOS	✓	53.4 (+0.8)	80.2 (+0.6)	65.2 (+8.1)
RetinaNet	✗	54.4	79.3	58.7
RetinaNet	✓	56.0 (+1.6)	80.4 (+1.1)	61.0 (+2.3)
Faster R-CNN	✗	53.5	81.4	58.2
Faster R-CNN	✓	57.8 (+4.3)	82.9 (+1.6)	64.7 (+6.5)
DETR	✗	52.1	76.8	54.9
DETR	✓	58.2 (+6.1)	81.1 (+4.3)	62.9 (+8.0)

Table 4. Transfer learning results on Pascal VOC benchmark. The prior knowledge learned by AlignDet pre-training can effectively help the downstream detection task.

different detection algorithms. AlignDet achieves a noticeable performance improvement in shorter training iterations and data settings. For example, AlignDet helps Mask R-CNN bring at least +3.2 mAP improvement in 12k iterations under diverse data protocols. Furthermore, the detectors suffer from the over-fitting issue in the low-data regime, in which Mask R-CNN with 90k iterations (15.2 mAP) is 3.9 mAP lower than with 12k iterations (19.1 mAP) at 5%

data. However, AlignDet still can obtain 1.4 mAP improvement under this dilemma, which means that the knowledge introduced by AlignDet pre-training can help the model avoid over-fitting. *Note that these non-trivial improvements are achieved by highly efficient pre-training. Only modules other than the backbone will be updated with 12 epochs pre-training on a relatively small COCO dataset, which is more efficient than other methods, as shown in Figure 2.*

Detection Backbones and SOTA model. As shown in Table 3, we validate the effectiveness of AlignDet on various backbones with the Mask R-CNN framework, including MobileNet v2 [51], ResNets [27], and Swin Transformers [40]. The experimental results show that our AlignDet can improve various backbones effectively. For example, AlignDet improves mAP by +1.2, +1.1, and +0.8 on supervised pre-trained MobileNet v2, ResNet-50, and Swin-Base, respectively. To further verify the effectiveness of AlignDet, we conducted advanced experiments on a mask image modeling pre-trained backbone (SimMIM [64] pre-trained Swin-Base), and the state-of-the-art detection model without additional detection data (CBNet v2 [34] with Swin-Large backbone). Both of them use strong data augmentation and train to convergence. The experiments demonstrate that AlignDet achieves consistent and impressive performance improvements on the SOTA model (CBNet v2) and advanced techniques (SimMIM).

Generalization Analysis. Table 4 shows the transfer results from COCO pre-training to PASCAL VOC fine-tuning. The VOC dataset has fewer categories compared to the COCO dataset, resulting in easier classification by the model. Therefore, we should focus more on the metric with a higher IoU threshold, *i.e.*, AP₇₅. The considerable improvement of AlignDet pre-training on AP₇₅ shows that our pre-training makes the predicted coordinates more accurate, which also reveals that the prior knowledge learned by AlignDet pre-training can be effectively transferred into downstream detection datasets and tasks.

Method	Dataset	Epochs	Mask R-CNN			RetinaNet		
			AP	AP ₅₀	AP ₇₅	AP	AP ₅₀	AP ₇₅
ReSim	ImageNet	400	40.3	60.6	44.2	-	-	-
InsLoc	ImageNet	400	42.0	62.3	45.8	-	-	-
SoCo	ImageNet	400	43.0	63.3	47.1	38.3	57.2	41.2
SoCo	COCO	530	40.6	61.1	44.4	-	-	-
Supervised	ImageNet	90	39.7	59.5	43.3	37.4	56.6	39.7
DetCo	ImageNet	200	40.7	60.5	44.6	38.0	57.0	40.5
DenseCL	ImageNet	200	40.3	59.9	44.3	37.5	56.0	39.8
DenseCL	COCO	800	39.6	59.3	43.3	-	-	-
Self-EMD	ImageNet	300	40.0	60.4	44.0	-	-	-
Self-EMD	COCO+	800	40.4	61.1	43.7	37.4	56.5	39.7
SlotCon	COCO	800	41.0	61.1	45.0	-	-	-
MoCo v2	ImageNet	800	40.3	59.9	43.9	37.8	56.4	40.4
+AlignDet	COCO	12	41.0	61.0	44.9	38.4	57.1	41.3
PixPro [†]	ImageNet	400	40.9	60.4	44.8	38.4	57.0	41.4
+AlignDet	COCO	12	41.7	61.7	45.5	39.0	57.5	41.9
SwAV	ImageNet	800	41.6	62.2	45.8	37.1	56.8	39.5
+AlignDet	COCO	12	42.3	62.5	46.7	38.5	58.2	41.0

Table 5. AlignDet achieves competitive results compared to other methods, with only 12 epochs pre-training on COCO. [†] denotes our reproduced results with officially released model and code.

Method	Backbone	Dataset	Epochs	AP	AP ₅₀	AP ₇₅
From Scratch	Sup. R50		-	39.5	60.3	41.4
From Scratch	SwAV R50		-	39.7	60.3	41.7
UP-DETR [12]	SwAV R50	ImageNet	60	40.5	60.8	42.6
DETReg [†] [2]	SwAV R50	COCO	50	39.8	61.0	41.6
AlignDet	Sup. R50	COCO	50	41.0	61.9	43.1
AlignDet	SwAV R50	COCO	50	41.4	62.1	43.8

Table 6. Comparasion of AlignDet with DETR-specific methods. AlignDet achieves SOTA performance with the same or fewer pre-training. All results are fine-tuned with 150 epochs on COCO.

Compare with Other Self-supervised Methods. As summarized in Table 5, we also validate the effectiveness of AlignDet on the self-supervised ResNet-50 backbones with Mask R-CNN and RetinaNet. All the experiments follow the 1x fine-tuning strategy defined in [24]. AlignDet can significantly improve the detector’s performance under various pre-trained backbones with only 12 epochs of box-domain pre-training on COCO. For example, AlignDet boosts the performance by +0.6 mAP under the RetinaNet with PixPro pre-trained backbone. AlignDet does not outperform SoCo on the Mask R-CNN because SoCo conducts end-to-end pre-training on most modules (except RPN) for a long time (400 epochs) on the full ImageNet (**5.3x larger than COCO**). When SoCo is pre-trained on COCO, its performance drops significantly. While AlignDet only needs 12 epochs (**33.3 \times acceleration**) on COCO to pre-train the modules out of the backbone. It owes to AlignDet decoupling the pre-training process and taking advantage of the existing pre-trained backbones to accelerate convergence. In addition, SoCo cannot adequately pre-train detectors decoupled from regression and classification branches (*e.g.*, RetinaNet), leading to sub-optimal performance. On the contrary, AlignDet achieves significant performance improvement with various detectors and backbones. We also compare with other self-supervised methods explicitly de-

Fine-tuning Schedule	AP ^{bb}	AP ₅₀ ^{bb}	AP ₇₅ ^{bb}	AP ^{mk}	AP ₅₀ ^{mk}	AP ₇₅ ^{mk}
1x	38.3	58.0	42.1	34.3	54.9	36.6
2x	38.8	58.4	42.4	34.7	55.4	37.2
3x	39.0	58.7	42.9	35.0	55.6	37.7
4x	39.2	59.5	42.9	35.6	56.5	38.1
1x (Ours)	39.4 (+1.1)	59.2 (+1.2)	43.2 (+1.1)	35.3 (+1.0)	56.1 (+1.2)	37.7 (+1.1)
2x (Ours)	39.8 (+1.0)	59.3 (+0.9)	43.3 (+0.9)	35.3 (+0.6)	56.4 (+1.0)	37.6 (+0.4)

Table 7. Results with longer fine-tuning epochs. Training the baseline (ImageNet pre-trained backbone) with longer epochs until convergence (4x) cannot bring similar improvements as AlignDet.

signed for DETR, as shown in Table 6. AlignDet shows noticeable improvements over these DETR-specific methods. For example, AlignDet with 50 epochs pre-training on COCO surpasses UP-DETR with 60 epochs on ImageNet by +1.1 mAP. *Notably, by incorporating the self-supervised backbone, we can pre-train all modules within various detectors in a completely unsupervised paradigm.*

Training Schedules. We conduct experiments with diverse training schedules to eliminate the concern that the performance gain from pre-training is due to longer training time. As shown in Table 7, the performance of AlignDet with 1x pre-training and fine-tuning is even 0.2 mAP higher than the 4 \times fine-tuning result of ImageNet pre-training, which indicates that the performance gain is mainly due to the design of AlignDet rather than the longer training time.

4.3. Ablation Study

We conduct a series of ablation studies to further understand the advantages of bridging the discrepancies between pre-training and fine-tuning in terms of data, model, and task. All experiments are conducted on Faster R-CNN with ResNet-50 [27] initialized by the supervised pre-trained weights. The pre-training and fine-tuning procedure adopts the standard 1x training schedule on COCO [37].

Data Discrepancy: Do data characteristics and domains matter in pre-training? Yes. Table 8 demonstrates the effectiveness of bridging the data discrepancy. Leveraging a single object-centric dataset such as ImageNet to pre-train the detector backbone will result in a poor performance of 37.9 mAP. By introducing box-domain pre-training in the COCO dataset, we can fully use multi-object datasets to mitigate inconsistencies in data characteristics, improving +1.1 mAP. To further verify the influence of the data domain, we also employ selective search to construct the ImageNet Subset, where the number of objects is consistent with COCO. There is still a noticeable performance gap between pre-training on ImageNet Subset and COCO, which shows that the data domain is essential in pre-training.

Model Discrepancy: Are all detector modules necessary to conduct pre-training? To answer this question, we

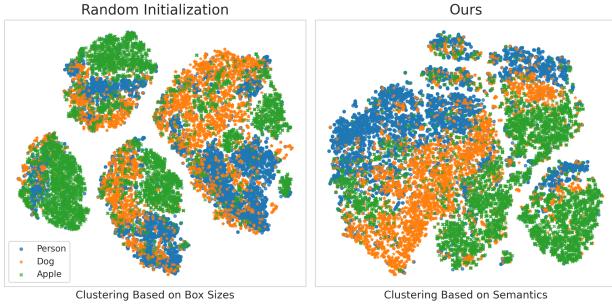


Figure 4. t-SNE visualization of ground truth annotations. AlignDet pre-training results in better class separation.

Image-domain	Box-domain	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
ImageNet	Random Initialization	37.9	58.2	41.1	22.0	41.4	48.6
ImageNet	ImageNet Subset	38.7	58.6	42.1	21.7	42.0	50.1
ImageNet	COCO	39.0	59.3	42.5	22.3	42.4	50.4

Table 8. Ablation study on data discrepancy. Both the data characteristics and domains matter in pre-training.

Backbone	Neck	RPN	Head	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
✓	✗	✗	✗	37.9	58.2	41.1	22.0	41.4	48.6
✓	✓	✗	✗	38.4	58.5	41.8	22.2	41.7	50.0
✓	✓	✓	✗	38.2	58.3	41.4	21.4	41.6	49.7
✓	✓	✗	✓	38.8	58.8	42.5	21.8	42.2	50.6
✓	✓	✓	✓	39.0	59.3	42.5	22.3	42.4	50.4

Table 9. Ablation study on the model discrepancy. Each module benefits from pre-training and leads to improvements.

load part of the pre-trained weights and remain other modules randomly initialization to investigate the benefit of pre-training different modules. As summarized in Table 9, we can obtain the following observations. i) Additional pre-training Neck improves the mAP by 0.5, of which the most significant improvement comes from AP₇₅ and AP_l. It indicates that Neck pre-training brings more accurate prediction boxes and larger size objects. ii) An interesting finding is that compared with the variants that the RPN and Head are all randomly initialized, merely pre-training the RPN causes performance decreases by 0.2 mAP. It may be because the randomly initialized head module cannot correctly handle the coarse results from the RPN. iii) On the contrary, if we pre-train other modules and maintain random initialization for RPN, the final performance is only 0.2 mAP lower than that of full pre-training. It may be because the RPN module only occupies 1.39% parameters within the whole detector. In conclusion, complete pre-training of the whole detector can bring the best downstream performance.

Task Discrepancy: Are detection-oriented pretext tasks helpful? As shown in Table 10, both the classification and regression pretext task can improve the detector accuracy, boosting the mAP from 37.9 to 38.7. Furthermore, the improvement brought by the classification pretext task is mainly reflected in the AP₅₀, while the regression task

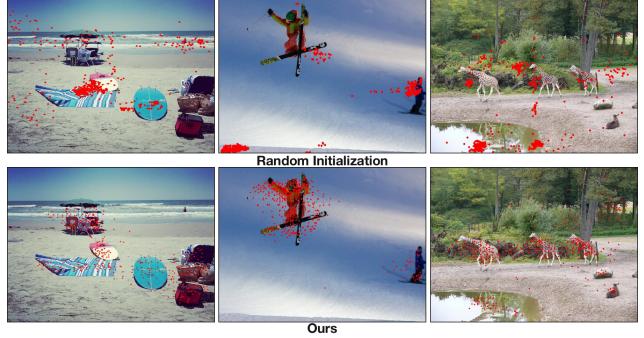


Figure 5. Visualization of predictions on COCO Val2017.

Image-domain	Box-domain	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l	
Classification	Classification	Regression						
✓	✗	✗	37.9	58.2	41.1	22.0	41.4	48.6
✓	✓	✗	38.7	58.8	42.1	22.4	41.9	50.2
✓	✗	✓	38.7	58.6	42.4	22.1	41.8	50.2
✓	✓	✓	39.0	59.3	42.5	22.3	42.4	50.4

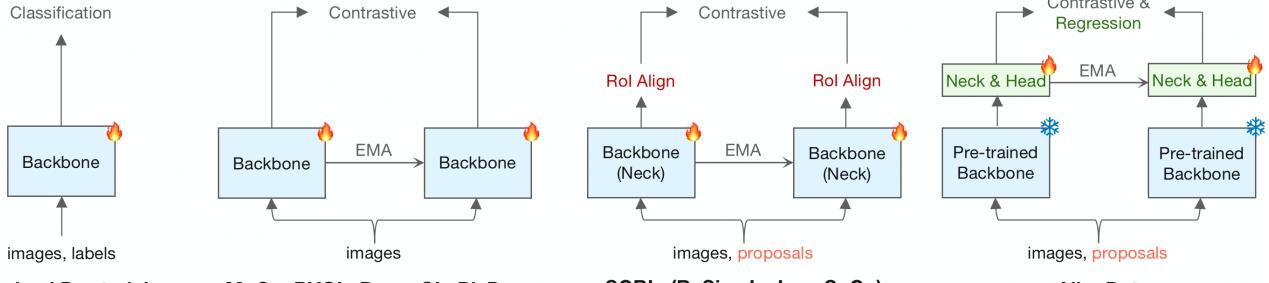
Table 10. Ablation study on task discrepancy. Both the classification and regression tasks are essential for AlignDet.

comes from the AP₇₅. It reveals that the classification pre-training brings higher prediction accuracy, while the regression pretext task helps to predict more accurate coordinates.

Effectiveness of Box-domain Pre-training. To visualize the effectiveness of box-domain pre-training, we depict t-SNE visualization [55] for features of 3 classes with each 5000 ground truth annotations. As illustrated in Figure 4, the features from the randomly initialized contrastive head f^{con} are clustered into four parts, with each cluster representing a feature level in the FPN. It reveals that the random initialization head classifies each box based on the box size. However, AlignDet provides a good weight initialization for f^{con} , leading to better class separation. We also visualize the center points of predicted boxes from Faster R-CNN in Figure 5. AlignDet focuses on potential objects instead of messy pixels compared to the random initialization results without box-domain pre-training. There are more visualization results of other methods in the supplementary material.

Freeze Backbone in Box-domain Pre-training. Freezing a layer or all parameters in the backbone is a common practice in previous work [26, 12, 2]. For example, Mask R-CNN freezes the first layer of ResNet when training on COCO, and UP-DETR and DETReg freeze the entire backbone during pre-training. Although there is no explanation for this approach in previous works, it significantly degrades the performance [12, 2]. Here we offer some conjectures as to why the backbone weights should be fixed during the box-domain pre-training stage, demonstrating that the frozen backbone leads to better performance.

The possible reason is that the amount of parameters is large for the detection model, while the data is relatively



Supervised Pre-training MoCo, BYOL, DenseCL, PixPro

Figure 6. Compared with other pre-training methods, AlignDet achieves adequate and more efficient pre-training for various detectors.

Frozen Backbone	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
✗	35.5	53.7	37.7	18.5	38.9	47.0
✓	37.3	56.6	40.1	21.0	40.9	49.8

Table 11. Ablation study on whether freezes the backbone during the box-domain pre-training.

Training Stage	Parameters	Dataset	Epochs	Training Time
Image-domain Pre-training	23.23 M	ImageNet	800	77.1 h
Box-domain Pre-training	8.79 M	COCO	12	5.6 h
Fine-tuning	32.02 M	COCO	12	4.4 h

Table 12. Ablation on training complexity for FCOS (ResNet-50)

small, and the label is noisy. If we train the entire detection model with small and noisy labels, the detector may overfit to these noise proposals, rather than learning semantic and position priors of objects. Therefore, in the box-domain pre-training, we freeze all parameters in the backbone. We also constructed experiments on Table 11 to demonstrate that fixing the backbone can bring more benefits.

Ablation on Training Complexity. As shown in Table 12, the box-domain pre-training introduces only a small computational overhead compared to the image-domain backbone pre-training (e.g., MoCo v2 here), yet achieves significant and wide-ranging improvements as shown in the paper, the training time is calculated on 8 NVIDIA A100 GPUs (80G). Please note we simply load the pre-trained backbone weights in practice, and both the selective search and our AlignDet are easy to implement.

5. Comparison with Other Methods

Compared with box-level contrastive learning counterparts (e.g., InsLoc [66], SoCo [58]), AlignDet makes full use of existing pre-trained backbones to enable efficient detection-oriented pre-training. We also show the pseudo-code of SoCo and AlignDet pre-training in the Supplement Material for a more intuitive comparison. In order to better understand the differences with other pre-training methods, we demonstrate the core pipelines of these methods in Figure 6. Compared with existing pre-training methods, AlignDet achieves adequate pre-training for all modules by introducing classification and regression knowledge, thus comprehensively addressing data, model, and task discrepancies between pre-training and fine-tuning.

6. Discussions

How do pre-training and fine-tuning attain alignment? AlignDet achieves alignments because the box-domain pre-training can be considered as a type of detection fine-tuning, the only differences are freezing the backbone and replacing the classification with box contrastive learning. To achieve *data alignment*, we utilize the same multi-object dataset to maintain uniformity in data properties and domain. The decoupled pre-training makes all modules can be well pre-trained to achieve *model alignment*. *Task alignment* is promoted via the integration of both regression and classification prior knowledge within the box-domain pre-training.

Is Selective Search Burdensome in Pre-training? No. The generation of selective search proposals is conducted offline and performed only once before pre-training. In box-domain pre-training, we randomly select the generated proposals as inputs, and operations of proposals do not bring too much overhead, as shown in Table 12.

Complexity of AlignDet. Although AlignDet requires three stages of image-domain pre-training, box-domain pre-training, and final fine-tuning. But we want to emphasize that both image-domain pre-training and fine-tuning are standard paradigms in object detection, so AlignDet only introduces an additional box-domain pre-training phase.

7. Conclusion

In this paper, we point out that there are data, model, and task discrepancies between pre-training and fine-tuning in object detection and propose a unified framework namely AlignDet to address these issues. AlignDet learns both classification and regression knowledge and enables highly efficient pre-training for all modules. Notably, it is the first framework to facilitate complete unsupervised pre-training of various detectors. Extensive experiments demonstrate that AlignDet significantly improves performance across diverse protocols. We believe AlignDet provides valuable insights and opens new avenues for visual pre-training.

References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021.
- [2] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *CVPR*, 2022.
- [3] Likun Cai, Zhi Zhang, Yi Zhu, Li Zhang, Mu Li, and Xiangyang Xue. Bigdetection: A large-scale benchmark for improved object detector pre-training. In *CVPR*, 2022.
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.
- [5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.
- [6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020.
- [7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.
- [9] Tom Tongjia Chen, Hongshan Yu, Zhengeng Yang, Ming Li, Zechuan Li, Jingwen Wang, Wei Miao, Wei Sun, and Chen Chen. First place solution to the cvpr’2023 aqtc challenge: A function-interaction centric approach with spatiotemporal visual-language alignment. *arXiv preprint arXiv:2306.13380*, 2023.
- [10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.
- [12] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *CVPR*, 2021.
- [13] Andong Deng, Xingjian Li, Zhibing Li, Di Hu, Chengzhong Xu, and Dejing Dou. Inadequately pre-trained models are better feature extractors. *arXiv preprint arXiv:2203.04668*, 2022.
- [14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.
- [16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.
- [17] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021.
- [18] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.
- [19] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.
- [20] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.
- [21] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Aistats*, 2011.
- [22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020.
- [23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.
- [24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.
- [25] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, 2019.
- [26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.
- [27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [28] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- [29] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.
- [30] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017.
- [31] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. *IJCV*, 2019.
- [32] Ming Li, Jie Wu, Jinhang Cai, Jie Qin, Yuxi Ren, Xuefeng Xiao, Min Zheng, Rui Wang, and Xin Pan. Parallel pre-trained transformers (ppt) for synthetic data-based instance segmentation. *arXiv preprint arXiv:2206.10845*, 2022.
- [33] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022.
- [34] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cb-net: A composite backbone network architecture for object detection. *TIP*, 2022.
- [35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

- [36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.
- [37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [38] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020.
- [39] W. Liu, Dragomir Anguelov, D. Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.
- [40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.
- [41] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 1982.
- [42] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings 5th Berkeley Symposium on Mathematical Statistics Problems*, 1967.
- [43] Matias Mendieta, Boran Han, Xingjian Shi, Yi Zhu, Chen Chen, and Mu Li. Gfm: Building geospatial foundation models via continual pretraining. *arXiv preprint arXiv:2302.04476*, 2023.
- [44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019.
- [45] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.
- [46] Jie Qin, Jie Wu, Ming Li, Xuefeng Xiao, Min Zheng, and Xingang Wang. Multi-granularity distillation scheme towards lightweight semi-supervised semantic segmentation. In *ECCV*, 2022.
- [47] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *CVPR*, 2023.
- [48] Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised pretraining improves self-supervised pretraining. In *WACV*, 2022.
- [49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015.
- [50] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *CVPR*, 2021.
- [51] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.
- [52] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019.
- [53] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019.
- [54] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 2013.
- [55] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.
- [56] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, 2021.
- [57] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.
- [58] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *NeurIPS*, 2021.
- [59] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. *NeurIPS*, 2022.
- [60] Jie Wu, Tianshui Chen, Hefeng Wu, Zhi Yang, Guangchun Luo, and Liang Lin. Fine-grained image captioning with global-local discriminative objective. *TMM*, 2020.
- [61] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *AAAI*, 2020.
- [62] Enze Xie, Jian Ding, Wenhui Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *CVPR*, 2021.
- [63] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, 2021.
- [64] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.
- [65] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *ICCV*, 2021.
- [66] Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. In *CVPR*, 2021.
- [67] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. *ICCV*, 2019.
- [68] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Kaiwen Cui, and Shijian Lu. Accelerating detr convergence via semantic-aligned matching. In *CVPR*, 2022.
- [69] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2022.

- [70] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. Dense siamese network. *ECCV*, 2022.
- [71] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *ICLR*, 2021.
- [72] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.
- [73] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, 2019.
- [74] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020.

A. Overview of Supplementary Material

The supplementary material is organized into the following sections:

- Section **B**: Implementation details for all experiments.
- Section **C**: More experiments and analysis.
- Section **D**: Visualization of the selective search proposals and the effectiveness of AlignDet.
- Section **E**: Broader impact and limitation.

B. Implementation Details

B.1. General Settings

Pre-training. All the hyper-parameters for box-domain pre-training follow the original fine-tuning settings except the prediction sampling procedure. For example, the learning rate is 2e-4 and weight decay is 1e-4 for Mask R-CNN [26] when fine-tuning on COCO [37] with ImageNet [14] pre-trained backbone. Hence in the box-domain pre-training stage, we set the same learning rate and weight decay to pre-train the modules out of the backbone. In terms of sampling predicted boxes, we select as many positive samples (predicted boxes that correspond to a ground truth proposal instead of background) as possible to expand the data for box-level contrastive learning. All methods apply the same pre-training data augmentation, which has been described in Section 4.1 of the main paper. All the experiments are pre-trained on the COCO train 2017 dataset with 12 epochs (1x), except 50 epochs for DETR [4]. During the pre-training stage, most experiments can be finished with 8 V100 GPUs (32 GB), which is efficient since we only train other modules out of the backbone (*i.e.*, neck and head).

Fine-tuning. During the fine-tuning stage, we use SyncBN [28] to calibrate magnitudes for pre-trained models following MoCo [24]. For the experiments with supervised pre-trained ResNet [27], we follow the default setting in mmdetection [7] to freeze the first layer of ResNet, and fine-tuning the other parameters under standard data augmentation with single-scale training. For the experiments with self-supervised backbones, we fine-tune all layers end-to-end with multi-scale training, and SyncBN is used across all layers, including the newly initialized batch normalization layers. For experiments with MobileNetv2 [51] and Swin Transformers [40], we follow the default training strategy defined in mmdetection. For the VOC [16] fine-tuning, we train 12k iterations to avoid over-fitting, and the learning rate is divided by 10 at $\frac{3}{4}$ and $\frac{11}{12}$ of total training time.

Since AlignDet pre-trains all modules in the detector and not just the backbone, we need to adjust the fine-tuned hyper-parameters to better transfer the pre-trained weights.

Thanks to the experience of previous work [58, 52, 33], adjusting the learning rate and weight decay is a good practice. The main principle of hyper-parameter adjustment in the fine-tuning stage is to increase the learning rate while reducing weight decay. The most common setting is to increase the learning rate by 1.5 times and reduce the weight decay to half of the original value. The specific values of different methods and experiments are listed in detail in each subsequent paragraph.

B.2. FCOS

FCOS [53] is a single-stage, point-based detector. The learning rate and weight decay are 0.1, 1e-4 for AlignDet pre-training and 0.15, 5e-5 for fine-tuning, respectively. The maximum number of sampled predicted boxes for the box-domain pre-training is 2048. Other hyper-parameters are set to the default values in mmdetection.

B.3. RetinaNet

RetinaNet [36] is a single-stage, anchor-based detector. The learning rate and weight decay are 0.1, 1e-4 for AlignDet pre-training and 0.15, 5e-5 for fine-tuning, respectively. The maximum number of sampled predicted boxes for the box-domain pre-training is 2048. Other hyper-parameters are set to the default values in mmdetection.

B.4. Faster R-CNN & Mask R-CNN

Faster R-CNN [19] and Mask R-CNN [26] are two-stage, anchor-based detectors. Here Faster R-CNN uses the ROI Align [26] operation. The maximum number of sampled predicted boxes for the box-domain pre-training is 4096. All the experiments including baseline results are re-implemented with the *4conv1fc* ROI head for a fair comparison, following previous work [25, 24]. For Faster R-CNN, we fine-tune with only object detection annotations, and for Mask R-CNN, we fine-tune with both object detection and instance segmentation annotations.

Specifically, for the supervised pre-trained MobileNet v2 and ResNet backbones, the learning rate and weight decay are 0.2, 1e-4 for AlignDet pre-training and 0.3, 5e-5 for fine-tuning, respectively. In our experiments, the weight decay should be smaller for the self-supervised ResNet-50 backbones, thus we set 5e-6 for PixPro [63] and MoCo v2 [10], and the learning rate is the same as pre-training, *i.e.*, 0.02. For SwAV [6] pre-trained backbone, the fine-tuning learning rate is 3e-2, weight decay is 5e-6, and warmup iterations are 1000. For Swin Transformer backbones, the learning rate is 1e-4 and weight decay is 5e-2 for AlignDet pre-training. During the fine-tuning stage, the learning rate is 1e-4 and weight decay is 2e-2.

Algorithm I SoCo Pseudocode, PyTorch-like

```
# x: input images
# p: selective search proposals
# aug: independent random augmentation

for x, p in data_loader:
    (x1, p1), (x2, p2) = aug(x, p), aug(x, p) # augmentation
    x1, x2 = backbone_q(x1), backbone_k(x2) # updated
    x1, x2 = neck_q(z1), neck_k(z2) # feature pyramid

    # proposals as final bboxes
    b1, b2 = p1, p2
    z1, z2 = roi_align(x1, p1), roi_align(x2, p2)
    z1, z2 = g_q(z1), g_k(z2) # feature projection

    L = loss_contrastive(z1, z2) # contrastive loss
    ema_update(backbone_q, backbone_k, neck_q, neck_k)
```

B.5. DETR

DETR [4] is a single-stage, query-based detector. A key factor that leads to slow convergence is the complication in aligning object queries with target features in different feature embedding spaces [68]. However, in the self-supervised setting, it is difficult to achieve this alignment because we do not have accurate semantic labels. To alleviate this issue, UP-DETR [12] initializes the query embedding with features extracted from cropped image patches. DETReg [2] predict the features of cropped image patches from the corresponding query embedding via L_1 loss. However, these approaches simply use foreground or background for bipartite matching under the unsupervised setting, lacking explicit semantic information for the label assignment. This paradigm leads to the mismatch between bipartite matching costs and loss calculation, which may cause unstable matching and affect the effectiveness of pre-training.

To address this challenge, we make a small modification to AlignDet. In addition to the common coordinate-based label assignment and contrastive learning, which is the same in other methods, we also introduce the category-based assignment and corresponding loss to pre-train DETR. Specifically, we crop the selective search [54] proposals from images and extract their features with supervised pre-trained backbones. Then we cluster the extracted features into 256 classes using the K-means algorithm [41, 42], the cluster results are regarded as pseudo-semantic labels to perform extra label assignment and cross-entropy loss to pre-train DETR. This has the advantage of introducing explicit category information into bipartite matching, which aligns label assignment and loss calculation in DETR, leading to more stable matching results. *Note that only DETR uses the clustering results of the features as extra pseudo-labels for box-domain pre-training, since the label assignment of other methods in this paper does not require explicit semantic information but only coordinates.*

We use both the default supervised pre-trained ResNet-50 [27] and the self-supervised pre-trained SwAV [6] for the experiments. The learning rate is 2e-4 for a batch size

Algorithm II AlignDet Pseudocode, PyTorch-like

```
# x: input images
# p: selective search proposals
# aug: independent random augmentation

for x, p in data_loader:
    (x1, p1), (x2, p2) = aug(x, p), aug(x, p) # augmentation
    x1, x2 = backbone(x1), backbone(x2) # frozen backbone
    x1, x2 = neck_q(x1), neck_k(x2) # feature pyramid

    # proposals as pseudo labels, boxes are predicted
    b1, b2 = head_q.f_reg(x1, p1), head_k.f_reg(x2, p2)
    z1, z2 = head_q.f_con(x1, b1), head_k.f_con(x2, b2)
    z1, z2 = g_q(z1), g_k(z2) # feature projection

    L = loss_con(z1, z2) + loss_reg(b1, b2, p1, p2) # losses
    ema_update(neck_q, neck_k, head_q, head_k)
```

of 64 during the box-domain pre-training stage, and the loss weights of contrastive loss and cross-entropy loss are 1.0. In the fine-tuning stage, the learning rate is 1e-4 for the batch size of 16, and we fine-tune all the parameters following previous work [12, 2]. Other hyper-parameters are set to the default values in mmdetection.

B.6. SimMIM and CBNet v2

To further verify the effectiveness of AlignDet, we conducted advanced experiments with mask image modeling pre-training method (SimMIM [64]) and SOTA detection algorithm (CBNet v2 [34]). We chose CBNet v2 because of its open source code and achieved SOTA performance without requiring additional training data (*e.g.* training on Objects365 [52]). However, since they do not open source the training code corresponding to the most powerful model, we use the officially released code, models, and configs to reproduce the results. More specifically, we use the **large scale jittering** [17] to fine-tune Mask R-CNN with 3x strategy (SimMIM pre-trained Swin-Large backbone), following the settings reported in the original paper. For CBNet v2, we use the **publicly released config** [34] to reproduce the results. Both external links are existing implementations that follow original papers, not part of our submission.

C. Further Analysis and Experiments

C.1. Pre-training with Longer Epochs

Pre-training the backbone for longer epochs does not necessarily lead to sustained performance improvements for downstream tasks, both for supervised [25] and self-supervised pre-training methods [58, 63, 24]. Here we find similar results on AlignDet, that is, 12 epochs pre-training is enough for AlignDet, as shown in Table 13 with RetinaNet. However, the pre-training for the backbone are usually hundreds of epochs. A potential reason for this phenomenon is that the pre-training parameters of the two are significantly different. In most object detection models, the number of parameters of the backbone is much more than that of the neck and head modules, so backbone pre-training often re-

Pre-training Schedule	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
1x	37.3	56.6	40.1	21.0	40.9	49.8
2x	37.0	56.2	39.3	20.8	40.6	49.5
3x	37.0	56.1	39.5	20.4	40.6	48.7

Table 13. Ablation study on pre-training schedules. All the results are fine-tuned with 12 epochs (1x schedule).

quires longer pre-training epochs to learn meaningful representation. On the contrary, since the neck and head modules have relatively few parameters, they can be well-trained with fewer epochs. Thus a longer pre-training time may lead to over-fitting and will not bring additional improvements. In addition, detection datasets such as COCO [37] are usually smaller than pre-training datasets (*e.g.*, ImageNet [14]), which may exacerbate this issue.

D. Visualization

D.1. Selective Search Proposals

We use the same selective search code and filtering strategy as SoCo [58] on the COCO train 2017 dataset, and apply non-maximum suppression (NMS) with a threshold of 0.5 at the end to remove redundant proposals. The images used for box-domain pre-training are shown in Figure 8.

D.2. Effectiveness of Box-domain Pre-training

Due to the significant differences in the design and mechanism of different detection methods, we need to design different visualization schemes to verify the effectiveness of our AlignDet under the unsupervised setting.

Faster R-CNN & Mask R-CNN. In this paper, the structural difference between Faster R-CNN and Mask R-CNN is only the presence or absence of a mask head, so they have the same prediction results and visualization for the detection task. In addition to the main paper, we also provide more visualizations here in Figure 9. Specifically, we use RPN to determine which of the predicted boxes are foregrounds and feed them into the head to get the predicted box coordinates. We plot the centers of these boxes instead of rectangles for better visualization. AlignDet focuses on objects instead of messy pixels compared to the random initialization results without box-domain pre-training.

Other Methods. Unlike Faster R-CNN or Mask R-CNN, other methods do not have an RPN module, which means we cannot determine which predicted boxes are foreground and which are background during inference. To demonstrate the effectiveness of AlignDet, we show the training losses in Figure 7 using RetinaNet as an example. AlignDet pre-training significantly accelerates the convergence of the model, with lower classification loss *loss_cls* and regression loss *loss_bbox* under the same training iterations.

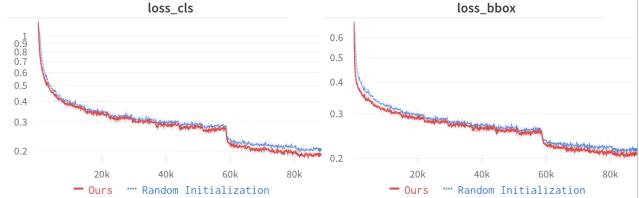


Figure 7. Fine-tuning losses of RetinaNet on COCO train 2017.

In addition, we also show the fine-tuning results of different detection models with or without AlignDet pre-training in Figure 10 to further demonstrate the effectiveness of our AlignDet. AlignDet achieves more accurate classification and precise coordinate results than the random initialization results without box-domain pre-training.

E. Broader Impact and Limitation

AlignDet represents a significant step forward in the development of unified and adequate unsupervised detection pre-training. Our approach enables the fully self-supervised pre-training of various object detection models, a milestone that was previously unattainable. Furthermore, the decoupled pre-training paradigm delivers highly efficient and effective pre-training, by separating the feature extraction from task-aware learning. **The decoupled pre-training paradigm can be readily extended to other vision tasks, allowing the integration of general-purpose pre-trained backbones with task-aware pre-trained necks and heads, which opens a door for solving the discrepancies between general pre-training and various downstream tasks.**

However, the dependence on selective search proposals in this paper may represent a potential limitation, we view it as a direction for future research. Overall, our work advances the state-of-the-art unsupervised detection pre-training and offers significant potential for improving the performance of object detection.

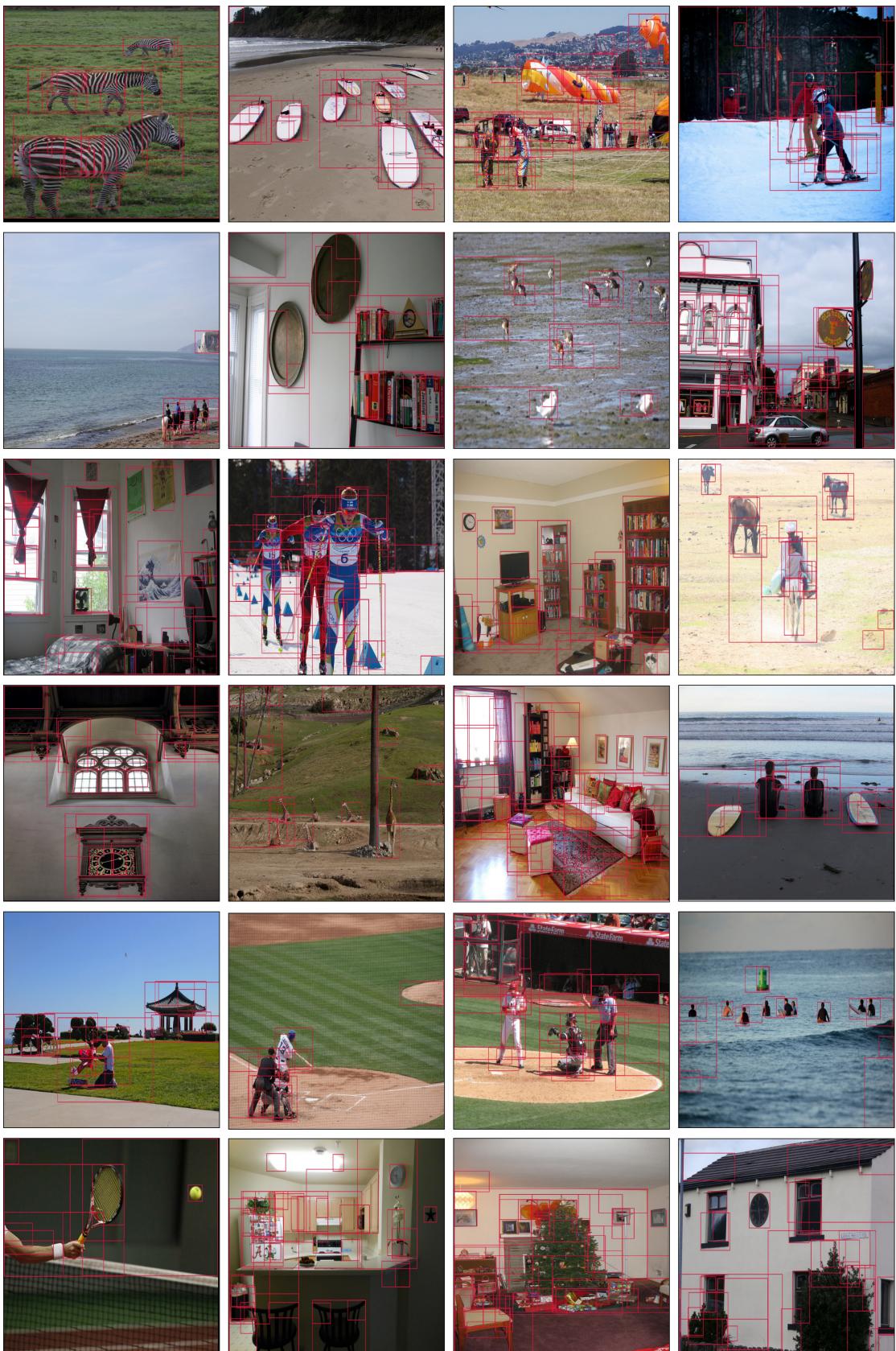


Figure 8. Selective search proposals on COCO train 2017 dataset.

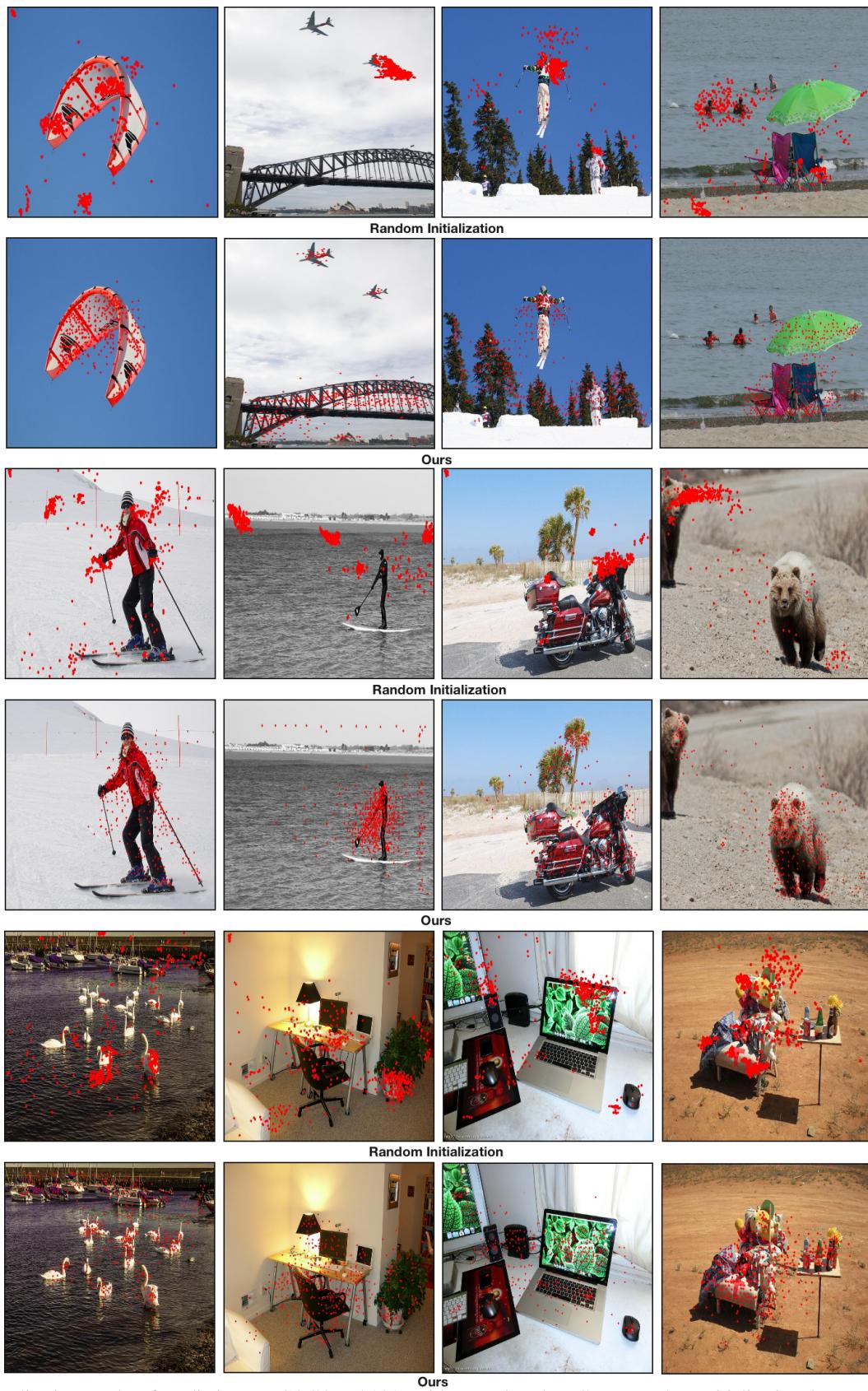
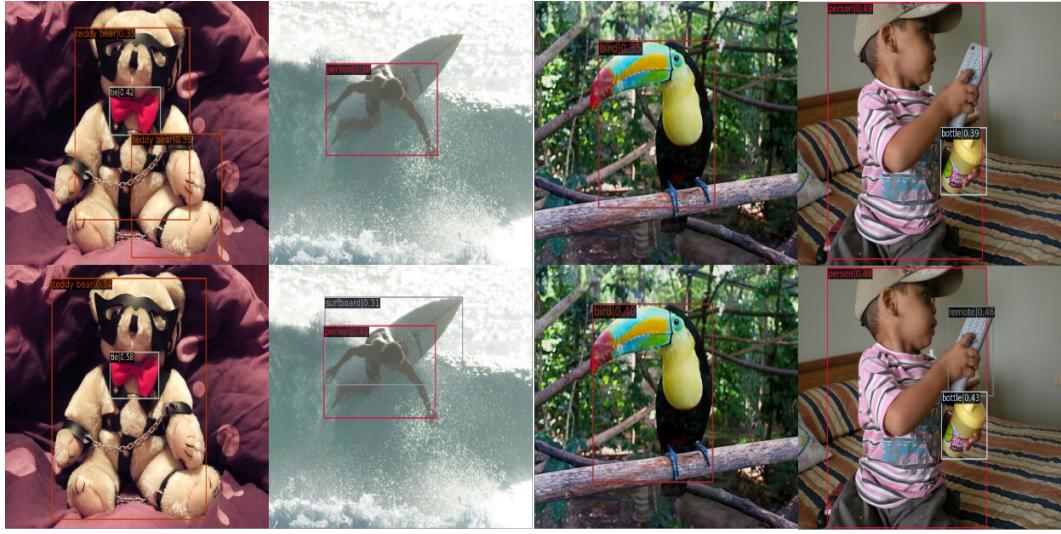


Figure 9. Visualization results of predictions on COCO Val 2017 with Faster/Mask R-CNN. Random Initialization denotes ImageNet pre-train, and ours means AlignDet pre-training.



Detection Results of FCOS



Detection Results of DETR



Detection Results of RetinaNet

Figure 10. Detection results with different models on the public COCO Val 2017 dataset. For each scene, the upper images are the fine-tuning results without box-domain pre-training, and the lower images are the fine-tuning results after the box-domain pre-training.