Integrally Migrating Pre-trained Transformer Encoder-decoders for Visual Object Detection

Feng Liu[1*] Xiaosong Zhang[1*] Zhiliang Peng[1] Zonghao Guo[1]
Fang Wan[1†] Xiangyang Ji[2] Qixiang Ye[1]
[1]University of Chinese Academy of Sciences [2]Tsinghua University
liufeng20@mails.ucas.ac.cn zhangxiaosong18@mails.ucas.ac.cn
pengzhiliang19@mails.ucas.ac.cn guozhonghao19@mails.ucas.ac.cn
wanfang@ucas.ac.cn xyji@tsinghua.edu.cn qxye@ucas.ac.cn

Abstract

Modern object detectors have taken the advantages of backbone networks pre-trained on large scale datasets. Except for the backbone networks, however, other components such as the detector head and the feature pyramid network (FPN) remain trained from scratch, which hinders the generalization capacity of detectors. In this study, we propose to integrally migrate pre-trained transformer encoder-decoders (imTED) to a detector, constructing a feature extraction path which is "fully pre-trained" so that detectors' generalization capacity is maximized. The essential differences between imTED with the baseline detector are twofold: (1) migrating the pre-trained transformer decoder to the detector head while removing the randomly initialized FPN from the feature extraction path; and (2) defining a multi-scale feature modulator (MFM) to enhance scale adaptability. Such designs not only reduce randomly initialized parameters significantly but also unify detector training with representation learning intendedly. Experiments on the MS COCO object detection dataset show that imTED consistently outperforms its counterparts by ∼2.4 AP. Without bells and whistles, imTED improves the state-of-the-art of few-shot object detection by up to 7.6 AP. Code is released at https://github.com/LiewFeng/imTED.

1. Introduction

Over the past two years, vision transformers (ViTs) [6] have been promising representation models. The vanilla transformer trained with a sophisticated self-supervised learning method, *e.g.*, masked autoencoder (MAE) [11], demonstrated great potential. Since the introduction of transformers [35] to computer vision, the effort of taming them for object detection has never stopped [3, 19]. This is motivated by the observation that ViTs pre-trained on extraordinarily large datasets incorporate over-completed and versatile features, which guarantee the performance and generalization capability of detectors finetuned on small datasets. [19, 17].

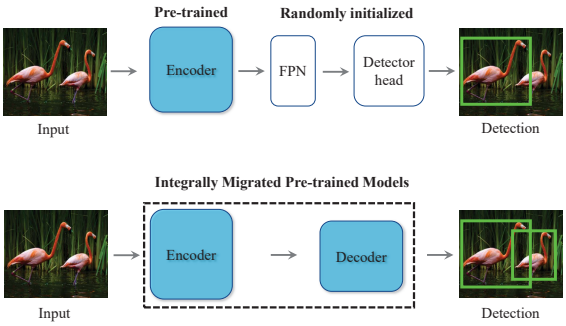


Figure 1: Comparison of the baseline detector *e.g.*, Faster R-CNN [30] using a transformer backbone (*upper*) with the proposed imTED (*lower*). The baseline detector solely transfers a pre-trained backbone network, *e.g.*, the encoder, but training the detector head and FPN from scratch. By contrast, our imTED approach integrally migrates the pre-trained transformer encoder-decoder. It significantly reduces the proportion of randomly initialized parameters and improves detector's generalization capability.

Modern object detectors, such as Faster R-CNN and Mask R-CNN [30, 12], typically consist of a backbone network, a neck component and a detector head. However, except for the backbone network, other components that occupy a significant proportion of parameters remain trained

*Equal Contribution.
†Corresponding Author.

from scratch, Fig. 1(upper). Such components, including but not limited to the region proposal network (RPN) [30], the feature pyramid network (FPN) [20] and the detector head [9], fail to take advantages of the representation models pre-trained on large-scale datasets.

In this study, we do not design any new components for object detection; instead, we devote to take full advantages of pre-trained models to improve detector's generalization capability. Specifically, we propose to integrally migrate pre-trained transformer encoder-decoders (imTED) to detectors, Fig. 1(lower), constructing a feature extraction path which is not only "fully pre-trained" but also consistent with pre-trained models, as much as possible.

As shown in Fig. 1(lower), imTED employs the ViT encoder pre-trained with MAE [11] as backbone, and uses the decoder as the detector head. It breaks the routine to remove the randomly initialized FPN from the feature extraction path while leveraging the adaptive respective field provided by the attention mechanism in ViTs [6, 28] to handle objects at multiple scales. These designs support the integral migration of pre-trained encoder-detector to the object detection pipeline. By adding linear output layers, $i.e.$, a light-weight classification layer and a bounding-box regression layer, atop the migrated encoder-decoder, imTED realizes object classification and localization. To enhance the capacity for multi-scale object detection, we introduce a multi-scale feature modulator (MFM), which combines both the advantages of FPN with those of fully pre-trained models.

The competitiveness of imTED is validated upon popular detectors including Faster R-CNN and Mask R-CNN [30, 12]. Experiments on the MS COCO dataset demonstrate that imTED with ViT-base model outperforms its counterpart by $\sim$2.4 AP at moderate computational cost. Benefiting from the integral migration of pre-trained models, imTED demonstrates strong generalization capability, which is validated by low/few-shot detection tasks. When reducing proportions of the training data, performance gains of imTED monotonously increase. When training a few-shot detector, by freezing the backbone network while finetuning the rest detector components, imTED improves the state-of-the-art by up to 7.6 AP. imTED opens up a promising direction for few-shot object detection using vision transformers.

The contributions of this study include:

- We integrally migrate pre-trained transformer encoder-decoders (imTED) to object detectors, constructing a "fully pre-trained" feature extraction path to improve detectors' generalization capacity.
- We redesign the feature extraction path to guarantee the "integral migration" of the pre-trained transformer encoder-decoders. We introduce a multi-scale feature modulator (MFM), to improve the scale adapatiblity of imTED.
- imTED not only achieves significant performance gains on object detection and few-shot object detection, but also takes a step towards unifying detector training with representation learning.

2. Related Work

Representation Models. Object detection has widely explored representation models pre-trained upon large-scale datasets. Over the past decade, CNNs [15, 31, 33, 13, 37] have been preferred representation models. Recently, vision transformers [6, 25, 7, 36] demonstrated greater potential. Vision transformers including ViT [6], Swin [25], MViT [7], and PvT [36] became promising models for image recognition. The vision transformers [1, 39, 27, 11] trained with self-supervised paradigms were validated to have higher generalization capability. Such generalization capability was pushed to a new height by MAE [11], which constructed not only representation models for feature extraction but also decoders for image reconstruction.

Model object detectors, either CNN-based [24, 23, 9, 30, 12] or transformer-based [2, 40], utilized pre-trained representation models as encoders to extract features, while left the FPN and detector head using randomly initialized parameters. These randomly initialized parameters, when finetuned using few training samples, experience difficult to achieve promising performance. Considering that the backbone, the FPN [22] and the detector head occupy most of the learnable parameters of an object detector, to make them be "fully pre-trained" is an important problem to be solved.

Feature Pyramid Network. FPN [22] leveraged a top-down structure with lateral connections to construct high-level semantic feature maps at scales, enhancing the flexibility for multi-scale representation. It was designed to adapt hierarchical CNN features but not compatible with plain representation models, $e.g.$, ViT [6]. To solve this problem, a small network was designed to obtain multi-scale features [19], but this unfortunately caused more parameters being randomly initialized.

The ViTDet method [17] proposed to remove the top-down feature fusion to simplify FPN, but remains not constructing a "fully pre-trained" feature extraction path. The major difference between ViTDet [17] and our imTED approach lies in the detector head. imTED simply feeds the last feature map of the MAE encoder to the RoI-Align component, without applying FPN. The aligned features are fed to the pre-trained transformer decoder for object classification and localization. Such designs guarantee that the feature extraction path be consistent with that of the pre-trained model.

Detector Head. DETRs [2, 40] are representative detectors, which leverage transformers as the detector head. Given CNN features as input, the transformer encoder-decoder reasons the relations of the objects and the global image context to output the final set of predictions. However, the vision transformers in DETRs were randomly initialized and only used to process features extracted by the backbone network. By contrast, the transformer in our imTED is pre-trained and utilized to not only extract features but also perform feature transformation. As a variant of DETR, ViDT [32] replaced the CNN backbone with a pre-trained transformer but still leaved the following transformer neck randomly initialized.

Recently, ViTDet [17] and MIMDet [8] tried the powerful representations pre-trained by MAE [11] for object detection. However, ViTDet solely leverages the pre-trained MAE encoder but deprecates the pre-trained decoder. Whereas, the proposed imTED utilizes both the pre-trained encoder and the pre-trained decoder. Although MIMDet [8] utilizes both the encoder and decoder for feature extraction, the core idea is leveraging the reconstruction ability of decoder to mask input image patches, which reduces the computation cost. It keeps the randomly initialized FPN and detector head, as well as introducing more randomly initialized layers for multi-scale feature extraction. By contrast, the imTED approach in this study utilizes the pre-trained encoder to extract features and the pre-trained decoder as the detector head, constructing a "fully pre-trained" feature extraction path, for the first time to our best knowledge.

3. Approach

The goal of this study is to integrally migrate the pre-trained transformer encoder-decoder as the pillars of an object detector. To this end, we choose encoder-decoders pre-trained by MAE [11] and migrate them to conventional two-stage detectors, *e.g.*, Faster R-CNN and Mask R-CNN [30, 12]. In what follows, we first describe the motivation of imTED. We then address how to integrally migrate the pre-trained encoder-decoders. Finally. we describe the implementation details of an imTED detector. We also show that modulating multi-scale features to the fully pre-trained feature extraction path further boosts the detection performance.

3.1. Motivation

MAE pre-trains encoder-decoder representation models based on the pretext task of masked image modeling [11]. By randomly masking image patches and reconstructing the masked patches, it trains an encoder for feature extraction and a decoder for image context modeling. It was validated that the MAE decoder has the ability to reconstruct masked pixels under a high mask ratio of 75% [11], demonstrating strong capacity to model image context information. This piques our curiosity: *could the spatial context modeling capacity of the MAE decoder benefits object localization?*

Table 1: Object detection and localization performance under three decoder variants on the ImageNet Localization Dataset. mAP and CoLoc are calculated under 0.5 IoU. CoLoc measures the correctly localized object ratio.

Model Variants	mAP	CoLoc	Acc.
pre-trained encoder	43.4	77.4	77.1
+ random decoder	43.9 (+0.5)	77.6 (+0.2)	77.7 (+0.6)
+ **pre-trained decoder**	44.8 (+**1.4**)	78.3 (+**0.9**)	78.0 (+**0.9**)

To answer this question, we conduct an experiment about single object detection on the ImageNet Localization Dataset [5][1] In the experiment, detectors are trained to predict a single object in each image to avoid the interference of complicated feature-object matching, design of FPN, and/or RoI alignment. Three variants of the object feature extractor are compared: (i) pre-trained encoder only; (2) pre-trained encoder with randomly initialized decoder; (3) pre-trained encoder with pre-trained decoder (imTED). Following the feature extractor, an object localization head and a classification head is used to realize object detection.

As shown in Table 1, the introduction of the randomly initialized decoder boosts the detection performance by 0.5 mAP and the localization performance by 0.2 CoLoc. Go a step further, the pre-trained decoder improves the detection performance by 1.4 mAP and the localization performance by 0.9 CoLoc. The significant performance gains validate that *the context modeling capacity of the pre-trained decoder does benefit object localization,* which motivates our integral migration approach.

3.2. Constructing A Fully Pre-trained Feature Extraction Path

Baseline Detectors. The Faster R-CNN [30] and Mask R-CNN [12] are employed as baseline detectors. The detector mainly consists of four components: a backbone network, a feature pyramid network (FPN), a region proposal network (RPN) and a detector head. By adding a mask head atop Faster R-CNN, Mask R-CNN can simultaneously conduct object detection and instance segmentation. The components of a conventional detector are partially pre-trained. The backbone network is pre-trained on large-scale datasets, while the FPN, RPN and detector head, which occupy a large proportion ($\sim 40\%$) of learnable parameters, are trained from scratch. The reason to use randomly initialized components lies in that the backbone networks specified for image classification [5] can not be directly applied

[1] Please refer to the supplementary material for details of dataset preparation.

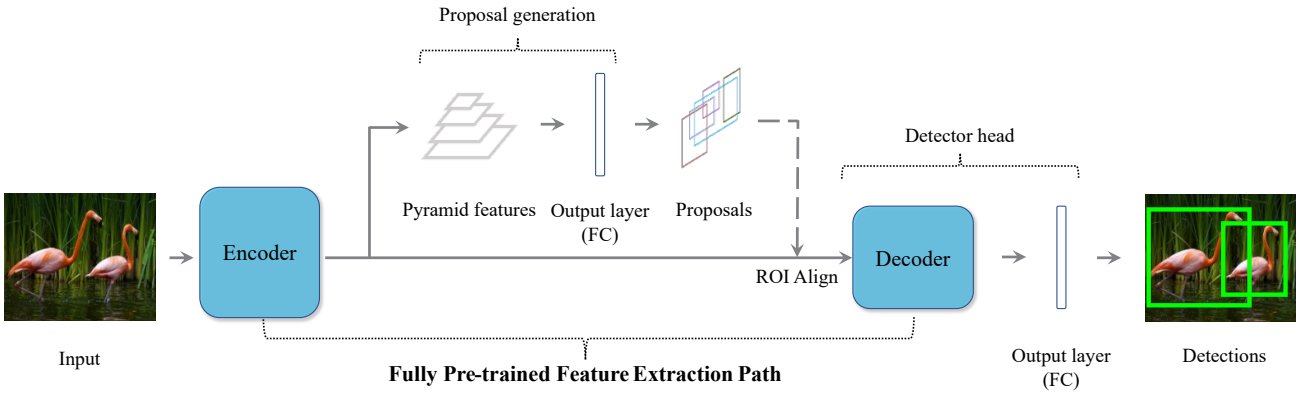


Figure 2: Architecture of the imTED detector. By integrally migrating the transformer encoder-decoder, imTED constructs a feature extraction path, which is "fully pre-trained". The reconstructed feature pyramid is only applied for object proposal generation but does not involve the feature extraction procedure. With these designs, the proportion of randomly initialized network parameters of the detector is significantly reduced.

for multi-scale feature extraction and object localization.

Integral Migration of Encoder-Decoder. As shown in Fig. 2, we redesign the feature extraction path by integrally migrating the transformer encoder and decoder pre-trained with MAE. The created imTED detector not only leverages the encoder for feature extraction but also the decoder for feature transformation. It then leverages a fully connected layer, a light-weight layer, for object classification and localization. Notice that the proposal generation pipeline remains unchanged, *i.e.*, the FPN and RPN remain using randomly initialized parameters. Whereas, the proposal generation pipeline is only responsible for producing region of interests (RoIs) but does not involve object feature extraction or transformation. Thereby, the randomly initialized parameters would not deteriorate detector's generalization capacity.

With these redesigns, the imTED detector has significantly fewer parameters trained from scratch, mostly lie in the proposal generation path, Fig. 2. When using the ViT-S [6] model, for example, the Faster R-CNN detector has $\sim$17.7M parameters trained from scratch, while imTED changes this figure to $\sim$3.3M, which infers a reduction of 81.3%. As is known, larger proportions of pre-trained parameters imply higher generalization capability. imTED thereby enjoys significantly higher performance than the baseline detector.

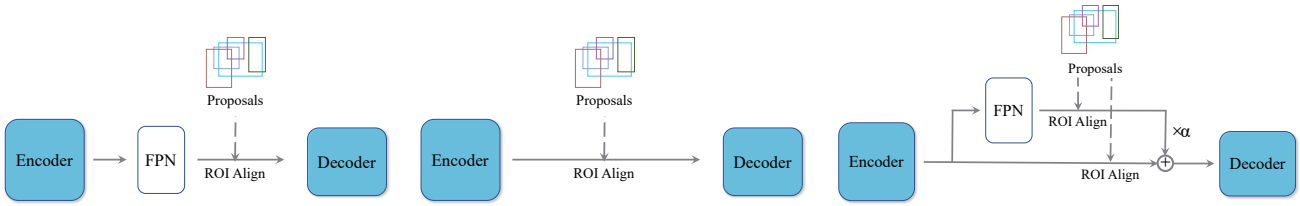
Removing Feature Pyramid Network. In Faster R-CNN, FPN can be deployed atop the encoder to augment the features to multiple resolutions Fig. 3(a). With FPN, large objects are represented by the low-resolution features and small ones by high-resolution features. However, FPN is constructed by using randomly initialized parameters, which violates the "fully pre-trained" idea. Fortunately, benefiting from the global attention mechanism, the transformer encoder is able to construct an adaptive receptive field [28], which reduces the requirement of scale alignment between objects and features. As a result, we are able to remove the FPN from the feature extraction path, Fig. 3(b). It is no doubt that removing FPN has a negative impact on the multi-resolution representation capability of features. Nevertheless, significant performance gains are observed in experiments, which supports the idea that constructing a "fully pre-trained" feature extraction path is more important than the multi-scale prior.

3.3. Detector Implementation

As described in Sec. 3.2, by migrating the transformer encoder as the backbone, plugging the decoder to the detector head, and removing the FPN, we construct a "fully pre-trained" feature extraction path. The architecture of RPN is not updated as it plays the role of generating region proposals but does not disturb the feature extraction stream. An imTED detector is then implemented by simply adding a few linear layers and a proposal generation module to the fully pre-trained encoder-decoder, Fig. 2.

Backbone Network. There is no modification to the transformer encoder except for resizing the encoder's positional embeddings so that they are consistent with input image sizes. The transformer encoder, pre-trained on a large-scale dataset, outputs a single-scale feature map which is down-sampled by a factor of 16 relative to the input image. The single-scale feature map is fed to the RoI-Align module for proposal feature extraction.

Region Proposal Generation. In the two-stage detection architecture [30], dense and multi-scale region proposals are used for object localization. To produce multi-scale feature maps, we up-sample or down-sample intermediate ViT feature maps by placing four resolution-modifying



(a) Faster R-CNN (b) imTED-SS (c) imTED

Figure 3: The involvement of (a) a conventional Faster R-CNN detector to (b) a single-scale detector (imTED-SS) and (c) the imTED detector with multi-scale feature modulating.

modules at equally spaced intervals of $d/4$ transformer blocks following [19], where d denotes the total number of blocks. The multi-scale feature maps are fed to the FPN, the output of which is further fed to the RPN for proposal generation. The training of the RPN parameters, $i.e.$, the weights of the fully connected output layer, is consistent with that of Faster R-CNN [30].

Detector Head. A pre-trained MAE decoder is migrated to the detector head to replace the randomly initialized network parameters, Fig. 2. The detector head consists of the pre-trained decoder and two linear layers. Given the feature map extracted by the encoder, an RoI-Align module extracts features for each region proposal. The extracted features are then embedded with location information by summarizing with position embeddings [11]. The features with position embedding are then fed to the decoder and transformed with alternative attention and MLP layers. The transformed features are finally fed to the linear classification and regression layers to predict object categories and location offsets.

3.4. Multi-scale Feature Modulator

Although the single-scale feature extracted by the transformer encoder is adaptive to object scales to some extent, we are wondering could the multi-scale feature representation be recalled back, in a new fashion, to further enhance the scale adaptability? To defend the idea of "fully pre-training", we can not directly call the FPN back to the feature extraction path; instead we redefine FPN as a multi-scale feature modulator (MFM), which acts after the RoI-Align module Fig. 3(c). Feature modulation for region proposals is defined as an adaptive linear weighting procedure, as

$$\mathbf{F} = \mathbf{F}_{ss} + \boldsymbol{\alpha} * \mathbf{F}_{ms}, \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ denotes the weighted features. $\mathbf{F}_{ss} \in \mathbb{R}^{C\times H\times W}$ denotes single-scale feature extracted by the pre-trained encoder and $\mathbf{F}_{ms} \in \mathbb{R}^{C\times H\times W}$ denotes multi-scale features extracted by the randomly initialized FPN, which is constructed by using the single-scale feature as input [20]. $\boldsymbol{\alpha} \in \mathbb{R}^{C}$ is a learnable weight vector. H and W respectively denote the height and weight of the output feature maps of the RoI Align module [30]. Both H and W are set to 7, following the setting of Faster R-CNN [30]. C is the channel dimension.

At the start-point of detector training, the elements of $\boldsymbol{\alpha}$ in Eq. 1 are initialized to zeros. When detector training proceeds, $\boldsymbol{\alpha}$ gradually updates so that the single-scale feature extracted by encoder is adaptively combined with the multi-scale features. In a learnable way, the multi-scale representation capacity is modulated to the single-scale representation. The evolution of Faster R-CNN to imTED-SS and imTED is illustrated in Fig. 3.

4. Experiment

4.1. Setting

The ViT models are categorized to ViT-S, ViT-B and ViT-L [6] according to the parameter scales. These models are pre-trained on ImageNet-1K using the self-supervised MAE method [11] for 1600 epochs. By adding a proposal generation module, RoI-Align module, multi-scale feature modulator and light-weight linear output layers atop the pre-trained encoder-decoder, the imTED detector is constructed. The detectors are evaluated on the MS COCO dataset [21], which consists of ∼118k training images and 5k validation images. Data augmentation strategies are defined by resizing image with shorter size between 480 and 800 while the longer side is no larger than 1333 [2] . The detector is trained using the AdamW optimizer [26] with a learning rate 1e-4, a weight decay of 0.05. The training lasts for 3×schedule (36 epochs with the learning rate decayed by 10 at epochs 27 and 33). The batch size is 16, distributed across 8 GPUs (2 images per GPU). For the ViT-S/B/L models, a layer-wise lr decay [1] of 0.75 and a drop path rate of 0.1/0.2/0.3 are also applied.

4.2. Detection Performance

In Table 2, imTED detectors are evaluated and compared with the baseline and state-of-the-art detectors. By replacing the ResNeXt101 backbone with a pre-trained ViT model, the baseline detector improves the average precision (AP) from 43.1 to 50.5, setting a solid baseline. Upon

Table 2: Object detection performance on the MS COCO dataset. Comparison of the proposed imTED detector with the state-of-the-art detectors using vision transformers as backbones. None of compared detection methods (ViTDet, MIMDet, imTED-SS and imTED) uses relative position embedding.

Approach	Backbone	Pre-train	Epochs	Mask R-CNN		Faster R-CNN		
				AP^{box}	AP^{mask}	AP	AP_{50}	AP_{75}
Baseline [37]	ResNeXt101	1k, sup	36	44.5	39.7	43.1	63.6	47.2
Baseline [25]	Swin-B	1k, sup	36	48.5	43.4	-	-	-
Baseline [18]	MViTv2-B	1k, sup	36	51.0	45.7	-	-	-
Baseline [30]	ViT-B	1k, MAE	36	51.3	45.3	50.5	71.4	55.5
ViT-Adapter [4]	ViT-S	1k, sup	36	48.2	**42.8**	-	-	-
imTED-SS(ours)	ViT-S	1k, MAE	36	48.0	42.4	47.3	68.6	51.0
imTED(ours)	ViT-S	1k, MAE	36	**48.7**	42.7	**48.2**	**68.4**	**52.6**
ViT-Adapter [4]	ViT-B	1k, sup	36	49.6	43.6	-	-	-
ViT-Adapter [4]	ViT-B	1k, MAE	50	50.8	45.1	-	-	-
Li et al. [19]	ViT-B	1k, MAE	100	50.3	44.9	-	-	-
ViTDet [17]	ViT-B	1k, MAE	100	51.6	45.9	-	-	-
MIMDet [8]	ViT-B	1k, MAE	36	51.7	46.1	-	-	-
imTED-SS(ours)	ViT-B	1k, MAE	36	52.3	46.0	52.2	72.8	57.1
imTED(ours)	ViT-B	1k, MAE	36	**53.3**	**46.4**	**52.9**	**73.2**	**57.9**
ViT-Adapter [4]	ViT-L	22k, sup	36	52.1	46.0	-	-	-
Li et al. [19]	ViT-L	1k, MAE	100	53.3	47.2	-	-	-
ViTDet [17]	ViT-L	1k, MAE	100	55.1	**48.9**	-	-	-
MIMDet [8]	ViT-L	1k, MAE	36	54.3	48.2	-	-	-
imTED(ours)	ViT-L	1k, MAE	36	**55.5**	48.1	**55.4**	**75.4**	**60.6**

Table 3: Ablation studies using ViT-S as the backbone (encoder) in 1x schedule. * indicates that the module is initialized using MAE pre-trained weights.

Detector Head	FPN	MFM	Params	FLOPs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Conv Layers	✓	✗	42.6M	403G	42.4	62.9	46.0	25.3	45.5	56.4
Decoder	✓	✗	30.1M	415G	42.2	62.4	45.8	25.8	45.0	57.0
Decoder*	✓	✗	30.1M	415G	42.5	63.0	46.1	25.6	45.4	57.7
Decoder*	✗	✗	30.1M	415G	43.2	63.9	46.9	25.0	46.6	58.6
Decoder*	✗	✓	30.3M	430G	44.0	64.6	47.6	26.2	47.3	59.3

the solid baseline, imTED-SS with ViT-B model improves the AP by 1.7 (from 50.5 to 52.2), which is a large margin for the challenging task. Note that this improvement is achieved without using FPN in the feature extraction path, which substantially validates the "integral migration" idea. When using multi-scale feature modulation (MFM), the total performance gain increases to 2.4 (52.9 *vs*. 50.5). imTED respectively improves the AP_{50} by 1.8 (from 71.4 to 73.2), and the AP_{75} by 2.4 (from 55.5 to 57.9). When using the Mask R-CNN framework, it respectively improves the AP^{box} by 2.0 and the AP^{mask} by 1.1, which are all significant margins. imTED also significantly outperforms the state-of-the-art detectors, $i.e.$, MIMDet and ViTDet, which use pre-trained transformers as backbones. ViTDet solely leverages the pre-trained encoder but deprecates the decoder. MIMDet leverages both the encoder and decoder for feature extraction but remains using a randomly initialized detector head, which deteriorates its generalization capability. imTED overcomes these disadvantages and achieves higher performance. Without using MFM, the AP^{box} and AP^{mask} of imgTED-SS respectively outperform the ViTDet detector (which uses FPN) by 0.7 (52.3 *vs*. 51.6) and 0.1 (46.0 *vs*. 45.9). When using MFM, the improvements of AP^{box} and AP^{mask} rise up to 1.7 and 0.5. When using the large backbone (ViT-L), the AP^{box} and AP^{mask} of imTED respectively outperform MIMDet by 1.1 and 0.9. Note that even only trained for 36 epochs, the imTED is comparable to, if not outperforms, ViTDet which is trained for 100

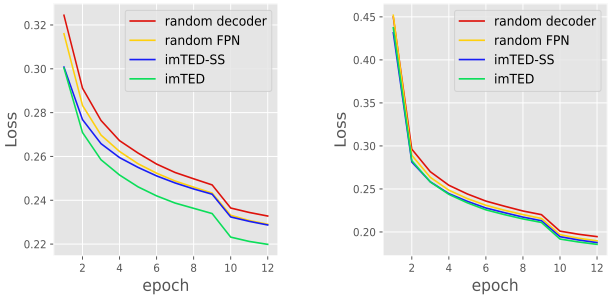


Figure 4: Comparison of object localization loss (left) and object classification loss (right).

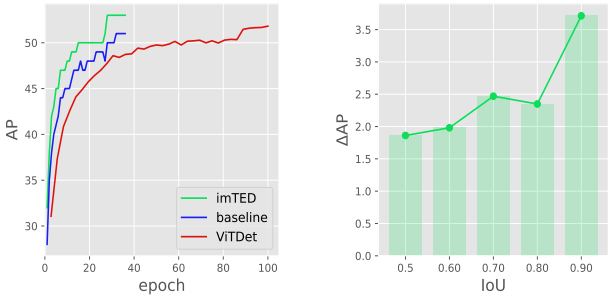


Figure 5: Performance gains during a 3× training with ViT-B. Left: AP improvements. Right: AP improvements under different IoU thresholds.

epochs.

4.3. Ablation Study

In ablations, we fine-tune the detector for 1× schedule (12 epochs with the learning rate decayed by 10× at epochs 9 and 11) on the *train*2017 split and evaluate on the *val*2017 split. By default, the ViT-S [34] is set as the backbone (encoder), and a 4-layer decoder with 256 dimensions is employed as the detector head. Unless otherwise specified, the ablation experiments are performed on Faster R-CNN.

Integral Migration. The baseline detector (Faster R-CNN) only uses a pre-trained encoder as backbone following [19]. Its predictions are obtained from FPN, convolutional (Conv) and fully connected layers in the detector head. By replacing the Conv layers in the detector head with the pre-trained MAE decoder and removing the FPN, we construct an integrally pre-trained feature extraction path. In Table 3, when replacing Conv layers in the detector head with a decoder without pre-training, there is a little performance drop -0.2 (42.2 *vs*. 42.4) observed. When using the encoder as the backbone and the decoder pre-trained by the MAE as the detector head, the AP performance improves 0.3 (42.5 *vs*. 42.2).

Removing FPN. By skipping FPN and constructing a fully pre-trained feature extraction path, imTED further improves AP by 0.7 (43.2 *vs*. 42.5). The total performance gain (43.2 *vs*. 42.4) over the baseline detector, considering the extensively investigated problem and the challenging aspects of the dataset, validates the effectiveness of the proposed imTED approach.

MFM. In Table 3, when using multi-scale features to modulate the feature extracted by fully pre-trained models, imTED improves AP performance by 0.8 (44.0 *vs*. 43.2). This shows the compatibility of fully pre-trained models with the randomly initialized module. In total, imTED improves the AP performance by 1.6 (44.0 *vs*. 42.4).

Training Loss Analysis. As shown in Fig. 4(left), imTED's localization loss decreases faster than the baseline detector using either a randomly initialized decoder or a randomly initialized FPN. imTED benefits from both the integral migration and multi-scale feature representation, demonstrating larger advantages on the localization ability. On the other hand, the compared detectors have similar classification loss curves, Fig. 4(right). This shows that imTED benefits a lot from the strong localization capacity of the pre-trained decoder.

Table 4: Detection performance using ViT-S in 1x schedule under different detector head depth.

Depth	FLOPs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1	371G	42.3	62.8	45.8	25.1	45.4	56.5
2	390G	43.1	63.4	46.8	26.1	46.3	57.4
3	410G	43.9	64.2	47.3	26.0	46.9	58.7
4	430G	44.0	64.6	47.6	26.2	47.3	59.3

Depth of Decoder. In Table 4, we evaluate the effect of depth of decoder (transformer blocks). Performance gradually saturates when the depth of decoder increases and the 4-layer decoder achieves the best performance. While larger objects benefit more from deeper decoders than smaller ones, the computational cost increases with the number of decoder layers.

Performance Gains During Training. imTED significantly improves AP during training, Fig. 5, which show that it not only speeds up training convergence but also raises the performance upper-bound. Particularly, imTED achieves larger performance gains under larger IoU thresholds, which implies improved localization capacity.

Computational Cost. In Table 3, replacing the Conv layers with the pre-trained decoder brings moderate increase of computational cost, *i.e.*, the FLOPs increases from 403G to 415G. When introducing MFM as the modulator, the FLOPs further increases from 415G to 430G. In total, the FLOPs increase by 6.7%.

4.4. Generalization Capacity

Low-shot Object Detection. imTED has greater generalization capacity because its feature extraction procedure is consistent with the pre-trained representation models. To validate this capacity, we evaluate the performance gains of imTED over the baseline detector by gradually reducing the training samples, which is termed low-shot object detection, Fig. 6(left). When the percentage of training data reduces, the performance gains of imTED over the baseline detector monotonously increase. Larger performance gains with less training data demonstrate greater generalization capability.

We also evaluate the detection performance of object categories under different numbers of training instances. As shown in Fig. 6(right), for the object categories of fewer training instances, imTED outperforms the baseline detector by larger margins. This further validates the effectiveness of imTED for low-shot object detection, which implies higher generalization capability.

Few-shot Object Detection. imTED can be applied for few-shot object detection without any modification. Following Meta YOLO [14], the object categories in MS COCO are divided into two groups: base classes with adequate annotations and novel classes with K-shot annotated instances. On MS COCO, 20 classes are selected as novel ones and the remaining 60 classes as base ones. The base classes are used to initialize the detector, $i.e.$, endowing it the ability to localize objects, through base training. The detector is then finetuned upon the novel classes for few-shot object detection. In Table 5, imTED respectively improves the state-of-the-arts of few-shot detection by 3.5 (19.0 to 22.5) and 7.6 (22.6 to 30.2) under 10-shot and 30-shot settings. The large performance gains further validate the generalization capability of the proposed imTED detectors.

Occluded Object Detection. We configure a sub-set of (534) images with occluded objects from the validation set of MS COCO. If two ground-truth objects has an IoU larger than 0.5, the corresponding image will be selected. In Table 6, by introducing the decoder and removing FPN, imTED improves the AP performance of occluded object detection by 1.2 (36.6 to 37.8). When using FPN as the modulator, the AP improvement increases to 2.4 (36.6 to 39.0). The total performance gain on the occluded subset is larger than that of the full set of MS COCO (2.4 *vs*. 1.6), demonstrating the superiority of integral migration on the occluded object detection task. As is known, MAE learns features via a form of denoising autoencoder, where each image is occluded with random patch masks and fed to the encoder while the decoder predicts the original pixel values of the masked (occluded) patches. This occlusion-and-prediction procedure performed on a large mount of images enables MAE models intrinsically learning occlusion invariant features. By integral migration, the imTED detector retains the capacity of MAE pre-trained models, which facilities detecting occluded objects.

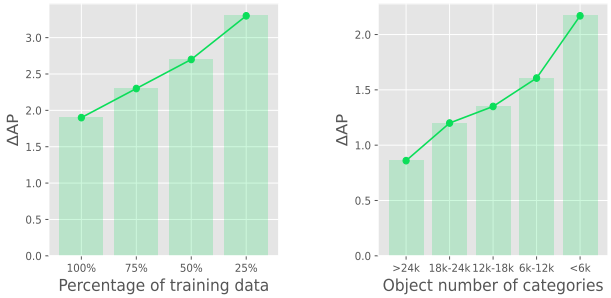


Figure 6: Performance gains on low-shot object detection. Left: Reducing training samples. Right: Training sample numbers.

Table 5: Performance comparison of few-shot object detection on the MS COCO dataset.

Shots	Method	Detector	AP
10	Meta YOLO [14]	YOLOv2	5.6
	CME [16]	FasterR-CNN + R101	15.1
	FCT [10]	PVTv2-B2-Li	17.1
	Meta-DETR [38]	DETR + R101	19.0
	DeFRCN [29]	FasterR-CNN + R101	18.5
	Baseline	Faster R-CNN + ViT-B	14.8
	imTED(ours)	imTED + ViT-B	**22.5**
30	Meta YOLO [14]	YOLOv2	9.1
	CME [16]	FasterR-CNN + R101	16.9
	FCT [10]	PVTv2-B2-Li	21.4
	Meta-DETR [38]	DETR + R101	22.2
	DeFRCN [29]	FasterR-CNN + R101	22.6
	Baseline	Faster R-CNN + ViT-B	22.2
	imTED(ours)	imTED + ViT-B	**30.2**

Table 6: Ablation studies using ViT-S as the backbone (encoder) on occluded objects in 1x schedule. * indicates that the module is initialized with MAE pre-trained weights.

Detector Head	FPN	MFM	AP	AP_{50}	AP_{75}
Conv Layers	✓	✗	36.6	55.5	38.8
Decoder	✓	✗	36.9	56.1	39.2
Decoder*	✓	✗	37.5	57.1	39.5
Decoder*	✗	✗	37.8	57.3	42.2
Decoder*	✗	✓	39.0	58.9	41.3

5. Conclusion and Future Remarks

We improved the conventional detection pipeline by integrally migrating pre-trained transformer encoder-decoders (imTED). The idea is to construct a feature extraction path which is not only "fully pre-trained" but also consistent with

MAE models. By migrating an MAE decoder to the detector head and removing FPN, imTED updated Faster R-CNN to a simpler yet more effective detector, where FPN can be optionally used as a feature modulator to further enhance scale adaptability. Experiments on low/few-shot and occluded object detection demonstrated the performance gains brought by imTED, with striking contrast with the state-of-the-arts. imTED provides an insight to fully exploit the potential of pre-trained masked autoencoders.

Despite the fact that imTED is implemented with less parameters, the computational cost of the decoder is moderately larger than the detector head with Conv layers. In the future, one solution is to use cascaded rejection strategies to reduce object proposals. The other solution is to configure a light-weight decoder using knowledge distillation. Another limitation of this work is that it's only applicable to the pre-trained models with both an encoder and a decoder.

Acknowledgement

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 62006216, 61836012, and 62225208, the Fundamental Research Funds for the Central Universities, and the Fundamental Research Funds for the Central Universities.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2, 5

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, volume 12346, pages 213–229, 2020. 2, 3, 5

[3] Wuyang Chen, Xianzhi Du, Fan Yang, Lucas Beyer, Xiaohua Zhai, Tsung-Yi Lin, Huizhong Chen, Jing Li, Xiaodan Song, Zhangyang Wang, et al. A simple single-scale vision transformer for object localization and instance segmentation. *arXiv preprint arXiv:2112.09747*, 2021. 1

[4] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 6

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255. IEEE CS, 2009. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2, 4, 5

[7] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *IEEE CVPR*, pages 6824–6835, 2021. 2

[8] Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, and Xinggang Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. *arXiv preprint arXiv:2204.02964*, 2022. 3, 6

[9] Ross Girshick. Fast r-cnn. In *IEEE ICCV*, pages 1440–1448, 2015. 2

[10] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with fully cross-transformer. *arXiv preprint arXiv:2203.15021*, 2022. 8

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 1, 2, 3, 5

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *IEEE ICCV*, 2017. 1, 2, 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 2

[14] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *IEEE ICCV*, pages 8420–8429, 2019. 8

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25, 2012. 2

[16] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond max-margin: Class margin equilibrium for few-shot object detection. In *IEEE CVPR*, pages 7363–7372, 2021. 8

[17] Yanghao Li, Mao Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 1, 2, 3, 6

[18] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. *arXiv preprint arXiv:2112.01526*, 2021. 6

[19] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 1, 2, 5, 6, 7

[20] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *IEEE CVPR*, pages 936–944, 2017. 2, 5

[21] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 5

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE CVPR*, pages 2117–2125, 2017. 2

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE ICCV*, pages 2980–2988, 2017. 2

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 2

[25] 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 *IEEE CVPR*, pages 10012–10022, 2021. 2, 6

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[27] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2

[28] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. *arXiv preprint arXiv:2105.03889*, 2021. 2, 4

[29] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *IEEE ICCV*, pages 8681–8690, 2021. 8

[30] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1, 2, 3, 4, 5, 6

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[32] Hwanjun Song, Deqing Sun, Sanghyuk Chun, Varun Jampani, Dongyoon Han, Byeongho Heo, Wonjae Kim, and Ming-Hsuan Yang. Vidt: An efficient and effective fully transformer-based object detector. *arXiv preprint arXiv:2110.03921*, 2021. 3

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE CVPR*, pages 1–9, 2015. 2

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, volume 139, pages 10347–10357, 2021. 7

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 1

[36] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE CVPR*, pages 568–578, 2021. 2

[37] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE CVPR*, pages 1492–1500, 2017. 2, 6

[38] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-detr: Few-shot object detection via unified image-level meta-learning. *arXiv preprint arXiv:2103.11731*, 2(6), 2021. 8

[39] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2

[40] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 3