FP-DETR: DETECTION TRANSFORMER ADVANCED BY FULLY PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale pre-training has proven to be effective for visual representation learning on downstream tasks, especially for improving robustness and generalization. However, the recently developed detection transformers only employ pre-training on its backbone while leaving the key component, i.e., a 12-layer transformer, being trained from scratch, which prevents the model from above benefits. This separated training paradigm is mainly caused by the discrepancy between the upstream and downstream tasks. To mitigate the issue, we propose FP-DETR, a new method that **F**ully **P**re-Trains an encoder-only transformer and smoothly fine-tunes it for object detection via a task adapter. Inspired by the success of textual prompts in NLP, we treat query positional embeddings as visual prompts to help the model attend to the target area (prompting) and recognize the object. To this end, we propose the task adapter which leverages self-attention to model the contextual relation between object query embedding. Experiments on the challenging COCO dataset demonstrate that our PT-DETR achieves competitive performance. Moreover, it enjoys better robustness to common corruptions and generalization to small-size datasets than state-of-the-art detection transformers. The source code will be made publicly available.

1 INTRODUCTION

Since the surge of deep learning-based object detector (Girshick et al., 2014), pre-training the backbone on large-scale datasets like ImageNet (Deng et al., 2009) and fine-tuning the model on downstream tasks has become a well-established paradigm. Pre-training significantly improves robustness (Hendrycks et al., 2019) and may enhance the model performance, especially on small datasets.

Following this paradigm, modern object detectors, like Faster RCNN (Ren et al., 2016) and YOLO (Redmon et al., 2016), generally add a *lightweight* task-specific layers on top of ImageNet pre-trained backbones, so that the model can enjoy the aforementioned benefits of pre-training. However, the recently developed detection transformers[1], e.g., DETR (Carion et al., 2020) and Deformable DETR (Zhu et al., 2020), only pre-trains its CNN backbone, while leaving the core module, *a 12-layer transformer* (including both encoder layers and decoder layers), trained from scratch in downstream tasks. As a result, the model suffers from limited robustness against common corruptions and the reliance on a large amount of training data for fine-tuning. Though UP-DETR (Dai et al., 2020) attempts to mitigate this problem by unsupervised pre-training, it requires an off-the-shelf CNN backbone that has already be pre-trained. Moreover, the two-stage separate pre-training methods may impair the model performance, since the two parts work together for object detection during fine-tuning.

The key reason that hinders existing detection transformers from benefiting from large-scale pre-training can be attributed to the discrepancy between the upstream task and the downstream task. Firstly, the object detection-oriented model structures are difficult to adapt to the ImageNet classification task. For example, the transformer decoder requires multiple query embeddings for detecting objects, while for ImageNet classification, there is only a single query embedding (class token) used. If the decoder is included during pre-training, both the self-attention layers and the projections on

[1]In this paper, we use detection transformer represents a class of object detectors built upon transformers, while DETR represents the seminal work by (Carion et al., 2020).

query embeddings in cross-attention layers may easily overfit to the single class token, making it is hard to pre-train the decoder. Secondly, the upstream ImageNet classification task misses some crucial components for the downstream object detection task. For example, while the downstream object detection task requires both localization and classification for the objects of interest, the upstream classification task only focuses on the latter. Moreover, the object relation modeling in object detection, which is important for removing heuristic post-processing like non-maximum suppression (NMS) (Carion et al., 2020), is absent in image classification.

To mitigate these problems, we propose FP-DETR, a novel method that reformulates the pre-training and fine-tuning phases for detection transformers. It fully pre-trains a detection transformer on ImageNet classification task and smoothly fine-tune it for object detection task though a task adaptor. Concretely, during pre-training, we introduce an encoder-only transformer structure by removing the decoder that can hardly be well pre-trained on the ImageNet classification task. Moreover, since both the CNN backbone and the transformer encoder in detection transformers can be seen as feature extractor, we replace the complex CNN backbone with a simple multi-scale tokenizer and only uses the transformer encoder for feature extraction. The resultant architecture is an efficient transformer encoder-only model. During fine-tuning on the object detection task, we take inspiration from the success of textual prompt (Liu et al., 2021a) in NLP and treat the query positional embeddings as visual prompts to help the model attend to the target areas (prompting) and recognize the object. Our key intuition is that if the model knows where to look at, it simply needs to identify the category of the object in hand, as did during pre-training. Based on this motivation, we devise a lightweight task adaptor to enhance the prompt ability of query positional embeddings and hence bridge the gap between upstream classification task and downstream object detection task. It leverages self-attention to emphasize the relation between query embeddings, such that the inter-object relationships, which have been neglected during pre-training, can be captured during fine-tuning.

Experiments show that the proposed FP-DETR achieves competitive performance on the challenging COCO dataset (Lin et al., 2014), and a better trade-off between the number of parameters and detection accuracy. Moreover, FP-DETR is more robust against common corruptions and can generalize well to small-size datasets like Cityscapes (Cordts et al., 2016), attributing to the effective fully pre-training.

2 Related Work

Object Detection. Object detection is one of the fundamental tasks in computer vision (CV). Representative object detection methods can be roughly categorized as the one-stage object detector, e.g., YOLO (Redmon et al., 2016), SSD (Liu et al., 2016), and two-stage object detectors like Faster-RCNN (Ren et al., 2016). While significant progress has been made, these methods are not end-to-end and rely on heuristic components like non-max suppression (NMS) and rule-based label assignments. Recently, DETR (Carion et al., 2020) formulates the object detection task as a set prediction problem and proposes an end-to-end pipeline to exploit transformer and bipartite matching for detection. The success of DETR brought the recent surge of detection transformers. For example, Deformable DETR (Zhu et al., 2020) proposes deformable attention to significantly accelerate the model convergence and reduce the computational cost, while allowing the model to benefit from multi-scale features. Conditional DETR (Meng et al., 2021) narrows down the spatial range for localizing the object regions via learning the decoder embedding conditioned on a spatial query. While effective, the crucial component in these models, i.e., a 12-layer transformer, is trained from scratch, which limits the robustness and generalization ability of the detection models.

Pre-training for Object Detection. The seminal work RCNN (Girshick et al., 2014) demonstrates the many benefits of ImageNet pre-training for object detection, e.g., better detection performance and faster convergence. Since then, pre-training and fine-tuning has become a well-established paradigm for object detection. Recently, He et.al challenge the common belief by showing that training from scratch on datasets like COCO with a longer schedule can also produce a competitive detection performance (He et al., 2019). However, Hendrycks et.al show that though comparable in-domain performance can be achieved, training from scratch results in much worse out-of-domain generalization (Hendrycks et al., 2019). Inspired by the success of pre-training for CNN-based object detectors, UP-DETR (Dai et al., 2020) proposes a novel random query patch detection pre-training task to improve DETR's performance and convergence. Though progress has been made,

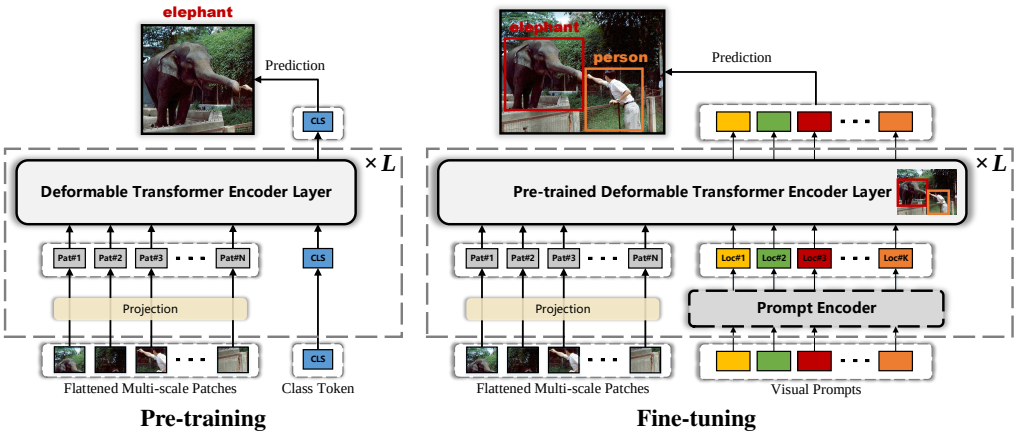

Figure 1: Illustration of pre-training and fine-tuning stages of FP-DETR.

their method relies on a pre-trained off-the-shelf CNN backbone. Moreover, the improvement is limited, since the CNN backbone and the transformer are pre-trained separately while they need to work jointly in the downstream task. Probably the most similar work to ours is YOLOS (Fang et al., 2021), which also pre-trains and fine-tunes an encoder-only transformer for object detection. However, it aims at demonstrating the feasibility of detecting objects using BERT (Devlin et al., 2018), with minimum inductive bias. As a result, the discrepancy between the upstream and downstream tasks, which significantly degenerates the model performance, is not considered in YOLOS. By contrast, we aim at maximizing the benefits of pre-training for the downstream object detection task. To this end, we treat query positional embeddings as visual prompts and devise a novel task adaptor to smoothly fine-tune the pre-trained model on downstream tasks. Moreover, since object detection task often requires high-resolution images as input, inductive bias like sparsity is introduced to reduce the computational costs.

Prompt-based Learning in NLP. Large-scale pre-trained models like GPT-3 (Brown et al., 2020) advance the field of NLP. However, such large-scale models are not designed for fine-tuning. To better utilize these models for downstream tasks, the "pre-train, prompt, and predict" paradigm emerges in the NLP community, and has drawn increasing attention (Liu et al., 2021a). The new paradigm reformulates the downstream task to mimic the task during pre-training with the help of a textual prompt, to bridge the gap between upstream and downstream tasks. Specifically, pioneer work introduces manually designed prompts to improve the fine-tuning (Schick & Schütze, 2020), or searches the textual prompt for downstream tasks (Gao et al., 2020; Jiang et al., 2020). Though improvements have been made, these prompts correspond to natural language phrases in discrete space can be sub-optimal, since the neural networks are inherently continuous. To tackle this problem, P-tuning (Liu et al., 2021b) proposes to optimize contiguous prompt embedding to better close the gap between pre-training and fine-tuning. In this paper, instead of reformulating the downstream task, we propose a new perspective and treat the query positional embeddings in detection transformers as visual prompts. To this end, a task adaptor is devised to facilitate fine-tuning on the downstream task.

3 Method

In this section, we first revisit the preliminaries on detection transformers, then introduce the pre-training and fine-tuning phases of our method as shown in Figure 1. Through exploiting the large-scale pre-training on ImageNet, our detection transformer can enjoy better detection performance, improved robustness, and enhanced generalization.

3.1 Preliminaries on Detection Transformers

Detection transformers generally consist of the following parts: a backbone, a transformer encoder, a transformer decoder, and a task-specific feed-forward network (FFN). The backbone F can be decomposed as multiple stages, i.e., $F = f^l \circ \cdots \circ f^2 \circ f^1$, where each stage f_i takes the output from the previous stage as input and outputs the down-sampled feature map at stage i. For an

input image $I \in \mathbb{R}^{H \times W \times 3}$, the backbone extracts multi-scale features, which are flattened and projected to 1-D sequence representation. Afterwards, the sequence representation is added with position and level embedding to obtained $x \in \mathbb{R}^{N \times D}$, where $N = \sum_{l=1}^{L} HW/S_l^2$ is the sequence length, S_l is the down-sampling rate at l-th feature level, and D is the dimension of the feature embedding. The transformer general consists of both encoder and decoder, and takes the sequence representation extracted from the image and the query embeddings as input for context modeling. The final prediction is made by applying FFN to the object representations.

While Deformable DETR (Zhu et al., 2020) contains four different feature levels, and the down-sampling rate ranges from 8, 16, 32, to 64, respectively, most other detection transformers, including DETR (Carion et al., 2020), Conditional DETR (Meng et al., 2021), contains only one feature level, whose down-sampling rate is 32.

3.2 PRE-TRAINING

Encoder-only Transformer. As described in Section 1, for image classification, only one class token is required for context aggregation. The six decoder layers trained on only one class token is prone to overfitting. To tackle this problem, we remove the decoder during pre-training on the image classification task, which leads to an encoder-only transformer structure. We follow Deformable DETR (Zhu et al., 2020) to adopt multi-scale deformable attention with four different feature levels to build our model. It largely alleviates the slow convergence of DETR (Carion et al., 2020), with the help of inductive bias like sparsity.

To perform image classification, an additional class token added with corresponding positional embedding $x_{cls} \in \mathbb{R}^D$ is concatenated with the sequence feature, for aggregating global context from the image tokens. Thus the input to the encoder-only transformer z_0 can be written as,

$$z_0 = \left[x_{cls}; x^1; x^2; \cdots; x^N\right]. \tag{1}$$

Afterward, the input sequence feature is processed by T different transformer encoder layers, each consists of a multi-scale deformable self-attention (MSDSA) layer and an MLP layer. Residual connections and layer norm (LN) are applied after each sub-layer, i.e.,

$$z'_t = \mathrm{LN}(\mathrm{MSDSA}(z_{t-1}) + z_{t-1}) \qquad t = 1 \ldots T, \tag{2}$$

$$z_t = \mathrm{LN}(\mathrm{MLP}(z'_t) + z'_t) \qquad t = 1 \ldots T. \tag{3}$$

Final predictions are made on class token at the T-th layer, i.e., z_T^0. During implementation, the class token is treated as an additional feature level, with only one resolution. Since multi-scale deformable attention assumes different levels of feature maps are spatially aligned, which does not hold for the class token in x_{cls}, attention mask is utilized to remove the information flow from x_{cls}.

Lightweight Multi-scale Tokenizer. The transformer encoder's ability to serve as a general feature extractor has been proved by recent advances in vision transformers (Dosovitskiy et al., 2020). While existing detection transformers contain both sophisticated backbone and transformer encoder for feature extraction, we argue that such design brings unnecessary complexity to the detection model. Instead, we replace the complex backbone with simple convolutional layers that work as a multi-scale tokenizer. Specifically, the feature extractor at each stage f_i is a simple single-layer convolutional layer that is designed only for down-sampling the feature map to expected resolution. Such design significantly simplifies the model and makes it easy to pre-train the detection model on ImageNet classification task. The detailed structure of our multi-scale tokenizer is shown in Appendix A.1.

Discussion. Most of the existing vision transformers adopt an encoder-only structure (Dosovitskiy et al., 2020; Liu et al., 2021c). By reformulating the detection transformer to an encoder-only structure, we solve the problem that the decoder is difficult to pre-train and simplify the model structure. Moreover, it allows us to take advantage of the existing experience in training vision transformers. In Appendix A.2, we explore pre-training on ImageNet using the encoder-decoder structure, however, the performance is significantly worse than that of the encoder-only structure.

3.3 FINE-TUNING

During fine-tuning, the single class token is replaced by N_q query content embeddings (Carion et al., 2020) $x_{obj} \in \mathbb{R}^{N_q \times D}$ for object detection. The content embeddings are also added with query

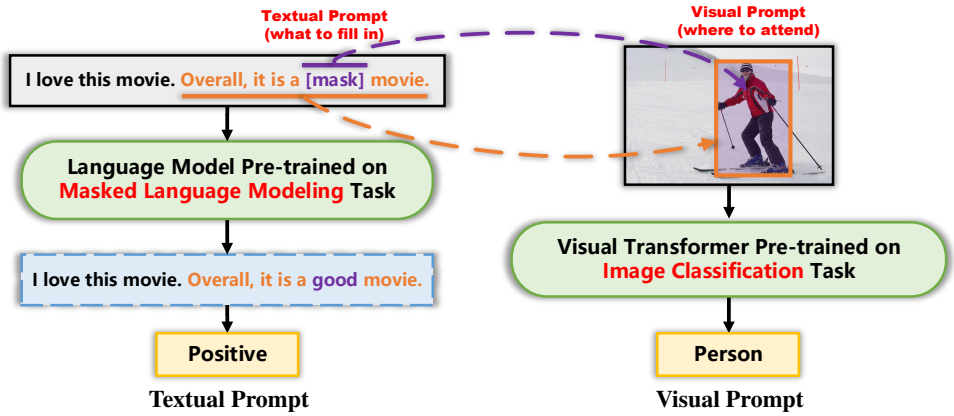

Figure 2: Analogy between textual prompt in NLP and visual prompt in CV.

positional embeddings (Carion et al., 2020) $p \in \mathbb{R}^{N_q \times D}$ before feed into the transformer encoder. Thus the input to the transformer encoder is:

$$\mathbf{z}_0 = \left[x_{obj}^1 + p^1; x_{obj}^2 + p^2; \cdots; x_{obj}^{N_q} + p^{N_q}; x^1; x^2; \cdots; x^N\right]. \tag{4}$$

The class token for image classification and the query content embeddings for object detection are both designed for context aggregation. While the former focuses on global context aggregation, the latter focus on aggregating the local context for one specific object instance.

Query Positional Embeddings as Visual Prompts. The textual prompt in NLP is shown in Figure 2. It reformulates the downstream task to mimic the pre-training one, so that the pre-trained model can better handle the downstream task. However, finding the optimal textual prompt is not easy, recent work such as P-tuning (Liu et al., 2021b) searches prompts in a contiguous space to bridge the gap between the upstream and downstream tasks.

Object detection can be decomposed into two subtasks, i.e., localization and classification. Our intuition is that if the classifier pre-trained on ImageNet knows where to look at, it can easily recognize the object within the specified region, as did during ImageNet classification pre-training. In our detection transformer, the query position embeddings are mapped to the reference points, which guide different query content embeddings to sample the corresponding context from the image content. In other words, the query position embedding works as a visual cue to point out the image region that the model should focus on, as shown in figure 2 (b). From this perspective, the query position embedding serves as a visual prompt, which is analogous to the textual prompt in NLP. Further, the process of training the model to localize objects corresponds to the process of searching textual prompts in contiguous space. And the process of final classification corresponds to the process of filling in the blanks and answers mapping in NLP (Liu et al., 2021a).

Improving Fine-tuning with Task Adaptor. Based on the above motivation, we propose a task adaptor to improve the query positional embeddings' ability to prompt. Specifically, the inter-object relationship can help the model better identify object regions. However, this crucial part is missing during ImageNet classification pre-training, since only one class token is used. To this end, we devise the task adapter to improve the modeling of the inter-object relationships. It pre-processes the visual prompts before they are sent to each pre-trained encoder layer, i.e.,

$$z_t^{0:N_q} = \text{TaskAdaptor}\left(z_t^{0:N_q}\right) \qquad t = 1 \ldots T. \tag{5}$$

By default, the task adaptor is instantiated as a self-attention layer. Afterward, the sequence is processed by the transformer encoder layers, as did in Equation 2 and 3. The object pattern embedding in the last encoder layer, i.e., $z_t^{0:N_q}$ is used for final prediction. Our task adaptor helps the visual prompts for individual objects to make associations with other instances, making them more suitable for the downstream object detection task.

Discussion. We propose a new perspective to view query positional embeddings as visual prompts. However, the existing pre-trained visual models are not as powerful as pre-trained language models (Brown et al., 2020). As a result, we are unable to fix the pre-trained models and only tune

Model	Layers	Hidden size D	MLP size	Heads	Params	Acc@Top-1
Lite	12	256	1024	8	11M	77.3
Small	12	384	1536	8	23M	80.8
Base	12	480	1920	10	35M	82.0

Table 1: Details of our FP-DETR variants. Parameters (Params) and top-1 accuracy (Acc@Top-1) here are calculated on the upstream ImageNet-1k (Deng et al., 2009) classification task.

the visual prompts, to perform few-shot or even zero-shot learning on downstream tasks, as did in NLP. But still, this new perspective helps us to better understand the detection transformers and the discrepancy between the pre-training and downstream tasks, which further helps us design the task adaptor to smoothly fine-tuning the pre-trained model on downstream tasks.

4 Experiments

4.1 Implementation Details

Datasets. Following the common practice, our detector is pre-trained on ImageNet (Deng et al., 2009) and fine-tuned on COCO 2017 (Lin et al., 2014) train set. Evaluation results on the val set of COCO 2017 are reported. To evaluate the models' robustness against common corruptions, we report the model performances on COCO-C (Michaelis et al., 2019), which is obtained by applying corruption synthesis algorithm to COCO. Besides, we evaluated the model's generalization ability by fine-tuning on the small-size dataset, i.e., Cityscapes dataset (Cordts et al., 2016).

Model Variants The lite version of our model follows the hyper-parameters of the Deformable Transformer in Deformable DETR (Zhu et al., 2020). Specifically, the encoder-only transformer contains 12 self-attention layers, each layer with 8 heads, the dimension of each head is 32, and the dimension of the MLP layer in FFN is 1024. Furthermore, we introduce two other model variants by changing the number of heads, or dimensions per head, while keeping the model depth unchanged. Table 1 summarizes the different variants of our model.

Training & Fine-tuning. Our FP-DETR is pre-trained on ImageNet-1k (Deng et al., 2009) for 300 epochs with AdamW (Loshchilov & Hutter, 2018) optimizer and cosine learning rate scheduler. Training strategies in DeiT (Touvron et al., 2021) are adopted, and the image size is set as 224×224. We use a batch size of 1,024 for training, and the initial learning rate is set as 5×10^{-4}. After pre-training, models are fine-tuned for 50 epochs with AdamW optimizer on the downstream tasks. The learning rate is initialized as 1×10^{-4} and decreased by a factor of 0.1 at the 40th epoch. We follow the implementation of Deformable DETR. Specifically, both the number of sampling points and the feature levels in multi-scale deformable attention are set as 4, and the number of object query embeddings is set as 300. Models are fine-tuned with a batch size of 32. All experiments are implemented on the NVIDIA A100 GPU.

4.2 Comparison with State-of-the-arts

As can be seen in Table 2, FP-DETR achieves competitive performance on COCO 2017 val set. Specifically, FP-DETR-Base is comparable with state-of-the-art Deformable DETR, while its small variant with 24M parameters can match the performance of UP-DETR and outperforms Conditional DETR and DETR with about 40M parameters. It performs better on detecting small objects and has better localization ability.

Moreover, FP-DETR shows a better trade-off between model parameters and detection accuracy compared to Deformable DETR. For example, FP-DETR-Small outperforms Deformable DETR with ResNet-18 (He et al., 2016) backbone, which has a similar amount of parameters. It also slightly outperforms Deformable DETR with ResNet-34 backbone, which has about 10M more parameters. The smallest version of our model, FP-DETR-lite, achieves 37.2 mAP on COCO 2017 with only 11M parameters.

Besides, FP-DETR significantly outperforms YOLOS-Base which also adopts a transformer encoder-only structure, in terms of both model efficiency and detection accuracy. This is mainly

Table 2: Comparision of FP-DETR with other detection transformers on COCO 2017 val set. Models are categorized as encoder-only (Enc) and encoder-decoder (Enc-Dec) according to the transformer structure. † indicates the model is pre-trained on ImageNet-21k.

Method	Structure	Backbone	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
DETR	Enc-Dec	ResNet-50	500	42.0	62.4	44.2	20.5	45.8	61.1	41M
UP-DETR	Enc-Dec	ResNet-50	300	42.8	63.0	45.3	20.8	47.1	61.7	41M
Conditional DETR	Enc-Dec	ResNet-50	50	40.9	61.8	43.3	20.8	44.6	59.2	44M
Deformdble DETR	Enc-Dec	ResNet-18	50	40.1	58.4	43.7	22.0	43.4	53.0	23M
Deformdble DETR	Enc-Dec	ResNet-34	50	42.3	60.7	46.0	24.2	45.8	56.1	33M
Deformable DETR	Enc-Dec	ResNet-50	50	43.8	62.6	47.7	26.4	47.1	58.0	40M
YOLOS-S	Enc	-	150	36.1	56.5	37.1	15.3	38.5	56.2	31M
YOLOS-B	Enc	-	150	42.0	62.2	44.5	19.5	45.3	62.1	127M
FP-DETR-Lite	Enc	-	50	37.2	56.5	40.4	21.7	40.0	48.6	11M
FP-DETR-Small	Enc	-	50	42.5	62.6	45.9	25.3	45.5	56.9	24M
FP-DETR-Base	Enc	-	50	43.3	63.9	47.7	27.5	46.1	57.0	36M
FP-DETR-Base†	Enc	-	50	43.7	64.1	47.8	26.5	46.7	58.2	36M

due to the inductive bias introduced in our model as Deformable DETR and the prompt-inspired task adaptor for more effective fine-tuning, which better bridges the gap between pre-training and fine-tuning.

The last row shows the result of our FP-DETR-Base fine-tuned from ImageNet-21 pre-training, denoted as FP-DETR-Base†. As can be seen, FP-DETR can benefit from pre-training at a larger scale, and FP-DETR-Base† with 36M parameters matches the performance of Deformable DETR with 40M parameters. In Appendix A.3, we also show the result of Deformable DETR with ResNet-50 backbone pre-trained on ImageNet-21k. Interestingly, the result is slightly worse than Deformable DETR with ImageNet-1k pre-trained backbone. We conjuncture that only pre-training part of the model makes it hard for the randomly initialized transformer adapts to the pre-trained CNN backbone, especially when the backbone has already been well-trained.

Table 3: Ablations on the task adaptor on COCO 2017 val set.

Task Adaptor	Shared	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
w/o	N/A	32.8	50.9	35.2	18.2	35.8	43.6	11M
Bi-LSTM		34.6	52.9	37.3	19.1	38.1	45.4	11M
Bi-LSTM	✓	34.9	53.0	37.5	20.2	38.0	46.9	11M
Self-attention		36.8	56.0	39.5	20.9	39.8	49.1	12M
Self-attention	✓	37.2	56.5	40.4	21.7	40.0	48.6	11M

4.3 Ablations on the Task Adaptor

To gain a better understanding of the effectiveness of the proposed task adaptor, we perform ablation studies on the lite version of FP-DETR, as shown in Table 3. We have the following observations. First, removing the task adaptor results in significant drops in model performance. Without the task adaptor, the model can not well capture the inter-object relationships, which is crucial for removing duplicates and improving object recognition. Second, the task adaptor instantiated with a bidirectional-LSTM layer also helps to adapt the pre-trained model to the downstream task, but it works slightly worse than a self-attention layer. We conjuncture that the bidirectional-LSTM is slightly worse on modeling long-range dependency compared to the self-attention layer, especially when the number of query embeddings reaches 300. By default, we adopt the self-attention layer for the task adaptor. In this way, our FP-DETR is a pure transformer encoder structure. Third, interestingly, a single task adaptor shared by visual prompts from different transformer encoder layers performs slightly better than specialized task adaptors for each layer. Since the task adaptor is trained from scratch on the downstream task, we conjuncture that the shared task adaptor sees diverse data from different levels compared to the non-shared ones, and is thus more sufficiently trained. These experiments validate the effectiveness of our task adaptor on modeling object relationships, which is crucial for object detection.

Table 4: Comparison of model robustness on the COCO-C dataset.

Method	Mean	Noise			Blur			
		Gauss	Shot	Impul	Defocus	Glass	Motion	Zoom
DETR	19.1	16.8	16.5	12.7	20.0	13.8	17.7	6.5
UP-DETR	21.6	22.1	22.2	18.7	20.1	13.4	18.3	6.9
Conditional DETR	20.3	18.6	18.6	16.0	20.5	14.8	18.4	7.1
Deformable-DETR	20.7	19.8	19.6	16.4	20.1	13.8	17.9	6.6
YOLOS-S	18.9	15.3	15.1	14.0	19.4	15.7	20.2	7.1
FP-DETR-lite	18.9	17.0	17.0	15.2	18.6	15.1	17.5	6.5
FP-DETR-Small	22.8	20.3	20.3	18.1	22.3	18.3	21.2	8.0
FP-DETR-Base	23.7	22.5	22.7	20.4	22.9	18.9	21.7	8.0
Method	Weather				Digital			
	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
DETR	16.4	21.4	29.0	35.0	18.8	23.2	18.3	19.8
UP-DETR	19.3	24.2	31.6	36.9	21.4	23.0	20.7	24.6
Conditional DETR	18.2	22.9	30.0	34.7	20.6	24.0	19.4	20.5
Deformable-DETR	18.3	23.6	31.7	37.0	21.2	24.0	19.3	21.0
YOLOS-S	17.7	21.2	26.5	29.5	19.7	22.8	17.1	21.8
FP-DETR-Lite	18.2	22.7	28.4	32.5	20.5	21.3	14.8	18.2
FP-DETR-Small	22.9	27.1	33.1	37.3	25.0	25.5	20.2	23.1
FP-DETR-Base	22.4	27.1	33.4	38.0	25.3	26.1	20.0	25.6

Table 5: Comparison of model generalization on the small-size Cityscapes dataset.

Method	Structure	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
DETR	Enc-Dec	500	17.3	38.0	14.1	3.4	14.9	37.1	41M
UP-DETR	Enc-Dec	300	22.0	43.0	19.2	3.3	17.3	44.3	41M
Conditional DETR	Enc-Dec	150	16.8	37.1	13.5	2.7	15.3	35.9	44M
Deformable DETR	Enc-Dec	50	29.9	53.3	29.5	11.7	30.8	48.5	40M
YOLOS-S	Enc	150	9.6	24.0	6.1	2.0	7.6	20.7	31M
FP-DETR-Lite	Enc	50	29.6	54.2	28.4	11.9	31.2	45.8	11M
FP-DETR-Small	Enc	50	31.4	55.7	30.9	11.9	33.5	48.8	24M
FP-DETR-Base	Enc	50	31.8	56.4	31.2	11.6	33.2	51.1	36M

4.4 Robustness to Common Corruptions

Model robustness is critical to trustworthy AI applications like autonomous driving. To this end, we evaluate the object detectors' robustness against common corruptions on COCO-C (Michaelis et al., 2019). As shown in Table 4, all detectors suffer significant performance drops under this rigorous condition. However, FP-DETR suffers the least performance drop compared to existing detection transformers. Notably, FP-DETR-Base performs best on 14 out of 15 types of corruptions, though it has comparable performance to Deformable DETR on the clean COCO 2017 val set. This is because the large-scale pre-training on ImageNet helps FP-DETR learn more generalizable representation. This is coherent with the observation in Hendrycks et al. (2019).

The first three rows in Figure 3 provide some qualitative results of our FP-DETR-Base on the COCO-C dataset. As can be seen, the images are significantly degenerated due to the existence of various corruptions, like shot noise, motion blur, fog, etc. However, FP-DETR-Base still manages to produce plausible results under low visibility and large distortion, which manifests its robustness.

4.5 Generalization to Small-size Dataset

In real-world applications, collecting a large amount of data is often infeasible. As a result, the models are required to perform well by training on only limited data. Table 5 shows the result of models fine-tuned on the Cityscapes dataset with only 2,975 training images. Though Conditional DETR can achieve competitive performance after training on COCO 2017 for 50 epochs, it fails to converge on the much smaller Cityscapes dataset with the same training schedule. Thus, we instead train it for 150 epochs. As can be seen, most detection transformers, including DETR, UP-DETR, Condi-

Figure 3: Qualitative detection results of FP-DETR on the COCO-C and Cityscapes datasets are shown in the first three rows and the last row, respectively.

tional DETR, and YOLOS-B, perform poorly under this condition. This is partly due to the lack of inductive bias in those models. As a result, the model requires a large amount of data to learn the sparsity that is beneficial for object detection on images. Deformable DETR performs better, thanks to the sparsity inductive bias introduced in the deformable transformer. However, it still performs worse than our FP-DETR, since the transformer in Deformable DETR is trained from scratch. Specifically, both FP-DETR-Base and FP-DETR-Small outperform Deformable DETR with more parameters. Notably, our FP-DETR-Lite with only 11M parameters matches the performance of 40M Deformable DETR, attributing to the end-to-end fully pre-training of the transformer.

The last row in Figure 3 shows some qualitative results of our FP-DETR-Base on the Cityscapes dataset. As can be seen, FP-DETR quickly adapts to the small-size dataset, and produces accurate object detection results, even for the small cars in the distant area. These results demonstrate the generalization ability of our method.

5 Conclusion

In this paper, we propose a novel detection transformer named FP-DETR that can take advantage of pre-training on large-scale upstream datasets to enable better robustness and generalization. FP-DETR contains a simple encoder-only structure for fully pre-training and can be smoothly fine-tuned for object detection via a task adapter. It leverages query positional embeddings as visual prompts to help the model attend to the target area (prompting) and recognize the object, effectively mitigating the gap between the upstream ImageNet classification task and the downstream object detection task. Experiments show that FP-DETR not only achieves competitive performance on the challenging COCO dataset, but also gains better robustness against common corruptions and generalization to small-size datasets.

ETHICS STATEMENT

We acknowledge that all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The authors strive to ensure the reproducibility of the experimental results. Specifically, implementation details are provided in Section 4.1, Section A.1, and Section A.3. The related studies with the same experimental setup are carefully cited and referred to. Moreover, the code will be made publicly available to ensure reproducibility.

REFERENCES

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *arXiv preprint arXiv:2106.00666*, 2021.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4927, 2019.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pp. 2712–2721. PMLR, 2019.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021a.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021b.

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. *arXiv preprint arXiv:2103.14030*, 2021c.

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

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE international conference on computer vision*, 2021.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*, 2020.

Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems*, 2017.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning and Representations*, 2020.

A APPENDIX

A.1 STRUCTURE OF THE MULTI-SCALE TOKENIZER

Our lightweight multi-scale tokenizer extracts feature maps of 4 different levels to construct the input sequences to the encoder-only transformer. Different from existing detection transformers, which use a sophisticated ResNet-50 backbone for feature extraction, our multi-scale tokenizer simply downsamples the input image to desired feature resolution. Specifically, we perform non-overlapping patch embedding, similar to the seminal ViT (Dosovitskiy et al., 2020). The only difference is that we perform multi-scale tokenization and adopt a hierarchical structure to progressively embed the tokens of large resolution. The detailed structure of our multi-scale tokenizer is shown in Table 6. The dimension D' is 192, 384, and 384, respectively for our Lite, Small, and Base model. After patch embedding, the 2D sequence is flattened to 1D sequence and further projected to tokens with dimension D.

Table 6: The architecture of the lightweight multi-scale tokenizer.

Multi-scale Tokenizer
Conv $8 \times 8 \times D'$, stride 8, pad 0
LayerNorm
Conv $2 \times 2 \times D'$, stride 2, pad 0
LayerNorm
Conv $2 \times 2 \times D'$, stride 2, pad 0
LayerNorm
Conv $2 \times 2 \times D'$, stride 2, pad 0
LayerNorm

A.2 EXPLORING ENCODER-DECODER STRUCTURE

While most detection transformers contain an encoder-decoder transformer, existing study on vision transformers extensively explores the encoder-only transformer for image classification. This results in a discrepancy between the upstream ImageNet classification task and the downstream object detection task. To mitigate this problem, we first explore pre-training the encoder-decoder transformer on the ImageNet classification task, using existing pre-training technologies (Touvron et al., 2021). Specifically, we pre-train and fine-tune an encoder-decoder transformer structure, denoted as FP-EncDec, on the ImageNet-1k dataset, follow the same implementation in 4.1. The model follows the design of our FP-DETR-Lite, except the last 6 layers of self-attention layers are replaced by 6 decoder layers and the task adaptor is removed. In each decoder layer, a standard self-attention layer (Vaswani et al., 2017) and a multi-scale deformable attention layer (Zhu et al., 2020) are applied sequentially, following Deformable DETR. Two variants of FP-EncDec are considered: (1) pre-training with a single class token, follow the common practice in training vision transformers (Dosovitskiy et al., 2020); and (2) pre-training with 300 class tokens, which equals the number of object query embeddings for fine-tuning. The 300 class tokens are pooled into a single class token before final classification.

Table 7: Comparison of encoder-decoder transformers and encoder-only transformer for pre-training and fine-tuning.

Method	Structure	Token Number	Acc@Top-1	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params
FP-EncDec	Enc-Dec	1	72.0	33.1	51.8	35.7	17.9	35.9	43.4	12M
FP-EncDec	Enc-Dec	300	76.4	35.2	54.0	37.9	20.9	38.1	47.0	12M
FP-DETR-Lite	Enc	1	77.3	37.2	56.5	40.4	21.7	40.0	48.6	11M

The results for both pre-training on ImageNet-1k and fine-tuning on COCO 2017 dataset are shown in Table 7. As can be seen, pre-training the transformer decoder with a single class token performs poorly on the ImageNet classification task, since both the self-attention layers and the projections on the class token in the cross-attention layer are trained on a single class token, which could easily

lead to overfitting. The low pre-training accuracy also limits the model's detection performance on the downstream task. By contrast, FP-EncDec pre-trained with 300 class tokens performs better on ImageNet-1k classification, the top-1 accuracy is 4.4 higher than the single class token counterpart, since the decoder is more sufficiently trained. However, it is still inferior to our encoder-only FP-DETR-Lite. We conjuncture that this is caused by (1) existing training techniques for vision transformers have been heavily tuned towards the encoder-only structure, which can be sub-optimal for the encoder-decoder transformer; and (2) the discrepancy between upstream and downstream tasks also degenerates the model's performance on object detection, even if the decoder has been pre-trained. We expect these findings may provide useful insights to the community to rethink the current paradigm of pre-training vision transformers, and pay more attention to the pre-training of encoder-decoder transformers.

A.3 Pre-training on ImageNet-21k

For ImageNet-21k pre-training, we follow the pipeline in Ridnik et al. (2021). The ImageNet-21k is first pre-processed by three steps: (1) cleaning invalid classes; (2) validation split; and (3) image resizing. Afterward, the model is trained on the processed dataset (ImageNet-21k-P) using the semantic softmax training (Ridnik et al., 2021). Specifically, our model is pre-trained on ImageNet-21k-P with a batch size of 4096 for 80 epochs. The model is initialized from ImageNet-1K pre-trained weights. The learning rate is initialized as 3e-4 and scheduled using one-cycle policy (Smith, 2018). RandAugment (Cubuk et al., 2020), Cutout (DeVries & Taylor, 2017), Label-smoothing (Szegedy et al., 2016), and True-weight-decay (Loshchilov & Hutter, 2018) are adopted for regularization. For more details on pre-training, please refer to Ridnik et al. (2021).

For fine-tuning FP-DETR further, we adopt the same implementation details in 4.1. For a fair comparison, we also finetuned Deformable DETR (Zhu et al., 2020) with an ImageNet-21k-P pre-trained ResNet-50 backbone. The ResNet-50 pre-trained weight is taken from the official release[2] of Ridnik et al. (2021).

Table 8: Comparision of models fine-tuned from ImageNet-1k and ImageNet-21k pre-trained weights. † indicates the model is pre-trained on ImageNet-21k.

Method	Structure	Bone	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params
Deformable DETR	Enc-Dec	R50	50	43.8	62.6	47.7	26.4	47.1	58.0	40M
Deformable DETR†	Enc-Dec	R50	50	42.9	62.1	46.6	25.1	46.3	57.7	40M
FP-DETR-Base	Enc	-	50	43.3	63.9	47.7	27.5	46.1	57.0	36M
FP-DETR-Base†	Enc	-	50	43.7	64.1	47.8	26.5	46.7	58.2	36M

The result is shown in Table 8. As can be seen, our FP-DETR can benefit from fully pre-training on the larger-scale ImageNet-21k dataset as well as fine-tuning with the task adaptor. Counter-intuitively, the performance of Deformable DETR degenerates when using the weights pre-trained on ImageNet-21k. We conjuncture that only pre-training the CNN backbone of the model makes it difficult for the randomly initialized transformer to adapt to the pre-trained backbone, especially when the backbone has already been well-trained. More research efforts should be made to further dig into this problem.

[2] https://github.com/Alibaba-MIIL/ImageNet21K