International Journal of Computer Vision

Detection Transformer with Multi-level Decoder

--Manuscript Draft--

Manuscript Number:	VISI-D-24-01852
Full Title:	Detection Transformer with Multi-level Decoder
Article Type:	Manuscript
Keywords:	Object Detection; Detection Transformer; computer vision; Multi-scale features; Transformer
Corresponding Author:	Yuan Ma Illinois Institute of Technology Chicago, IL UNITED STATES OF AMERICA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Illinois Institute of Technology
Corresponding Author's Secondary Institution:	
First Author:	Yuan Ma
First Author Secondary Information:	
Order of Authors:	Yuan Ma
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	In this work, we present a novel structure for the Detection Transformer (DETR) that efficiently introduces multi-scale features into the decoder. We found that the traditional design of the DETR decoder is inefficient due to its repetitive structure, leading to suboptimal performance. To address this issue, we leverage multi-scale features, which capture different levels of semantic information from the input image, and integrate them into various layers of the decoder. This design improves the model's efficiency by providing the decoder with richer information, enabling more accurate object detection. As a result, our model achieves over a 3% improvement compared to its state-of-the-art counterpart, DN-DETR, with almost no additional computational cost. Moreover, our DC5 model outperforms the Deformable version of DN-DETR, demonstrating that our proposed structure could surpass the Deformable architecture that has dominated DETR for years.

DETR with multi-level decoder

Yuan Ma

March 2024

Abstract

In this work, we present a novel structure for the Detection Transformer (DETR) that efficiently introduces multi-scale features into the decoder. We found that the traditional design of the DETR decoder is inefficient due to its repetitive structure, leading to sub-optimal performance. To address this issue, we leverage multi-scale features, which capture different levels of semantic information from the input image, and integrate them into various layers of the decoder. This design improves the model's efficiency by providing the decoder with richer information, enabling more accurate object detection. As a result, our model achieves over a 3% improvement compared to its state-of-the-art counterpart, DN-DETR, with almost no additional computational cost. Moreover, our DC5 model outperforms the Deformable version of DN-DETR, demonstrating that our proposed structure could surpass the Deformable architecture that has dominated DETR for years.

1 Introduction

In the past few years, object detection has evolved from traditional methods based on complex hand-crafted features to more advanced deep learning detector. However, popular deep learning object detection frameworks such as R-CNN [3] and its variants [5, 6, 17], still require complex multi-stage pipelines which heavily rely on handcraft features. Detection Transformer [2], known as DETR, emerged as a revolutionary approach that simplify the detection pipeline by leveraging the transformer architecture, which is traditionally used in natural language processing. Unlike other detection models, DETR enables end-to-end object detection without complex hand-designed components. This new approach quickly gained attention in the research community due to its elegance and efficiency.

However, the baseline DETR is notorious for its low accuracy on small objects and unacceptable long training schedule. Additionally, it is computationally expensive for DETR to process high-resolution images. In response to these challenges, Deformable DETR [12] has introduced deformable multi-head attention. This efficient adaptation enhanced the model's ability to process high-resolution images. Moreover, it proposed multi-scale deformable attention that

extends attention mechanism to multi-scale, improving the detection accuracy of small object. This multi-scale deformable attention has dominated multi-scale DETR for years and it is indispensable for DETR to attain state-of-the-art performance, ensuring that the end-to-end detector surpasses all its counterparts. However, deformable attention is not without its flaws. It compromises the model's ability to capture global information, consequently diminishing the accuracy in detecting large objects.

On the other hand, a series of works focus on optimizing object queries[3, 14, 17, 20, 21]. Object queries interfere with image features to predict position of the object, which are tend to be multi-pattern as shown in [14]. The multi-pattern behavior of object queries is harmful for the model to converge during training. Thus researchers provide prior information such as anchor point or anchor box to object queries, achieving faster convergence and better accuracy. While the prior information for object query solve the multi-pattern issue, few may notice that the object queries interact with the same feature map in repetitive decoder layers is also problematic. It introduces redundant information, reducing efficiency and resulting in a sub-optimal model. As illustrated in Fig.1, the accuracy of the original DETR shows minimal accuracy drop when the last two decoder layers are cut, with most performance coming from the first three layers. The same trend can also be found at DAB-DETR even with a careful designed decoder.

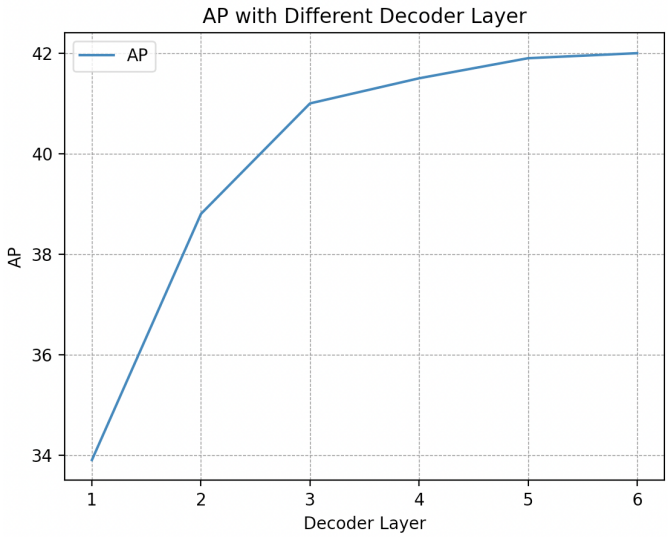


Figure 1: Average precison of DETR with different decoder layer

Motivated by this observation, we propose the DETR with Multi-level Decoder (MLD-DETR). The proposed multi-level design leverages multi-scale features different from any other multi-scale DETR (as shown in Fig.2(a) and Fig.2(b)), in the meanwhile naturally aligning with the function of dynamic anchor box proposed in DAB-DETR. Intuitively, to dynamically locate an object, one must

first form an initial hypothesis based on overview of the image, and locating the object by identification of specific details. In other words, we are detecting the object suceesively with the help of different semantic information. Based on this insight, our proposed multi-level decoder arranges the multi-scale feature maps into different decoder layers instead of feeding them into the decoder at once. Then the object queries are paired with feature maps containing different semantic representations in each layer, respectively. Our proposed method is simple and effective: with only 3 additional fully connected layers, we observed a 3.1% mean average precision increase on COCO2017 dataset compared to our baseline model (DN-DETR), with almost no extra computational cost. And our MLD-DETR-DC5 model achieves comparable accuracy with DN-deformable-DETR with less parameters and better detection accuracy on large objects.

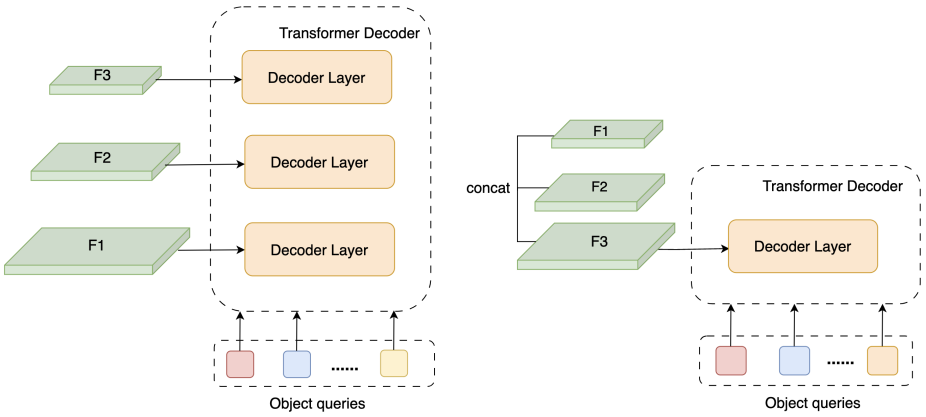


Figure 2: Comparision between our proposed multi-scale method and traditional multi-scale method in DETR. (a) shows ours multi-level design, the feature maps F1, F2 and F3 are resized to same resolution and flattened. (b) is multi-scale design in other DETR family. Note that feature maps F1, F2 and F3 are flattened in implement.

In summary, we describe our contributions as follows: 1. we rearrange various feature maps into different decoder layers in DETR, solving the inefficiency issues. 2. we propose a novel multi-level DETR pipeline with a unique arrangement of feature maps, offering an better multi-scale methodology than deformable DETR. 3. With proposed DETR with Multi-Level Decoder(MLD-DETR), our model can achieve better accuracy as well as faster convergence with almost zero additional computation cost compared to the baseline.

2 Related Works

Prior information for object query DETR is notorious for its slow convergence and low accuracy on small objects. Lots of researchers believe it is because of the poor behavior of object query during training. Consequently, numerous works have been proposed to modify DETR's query design[3, 14, 17, 20, 21], aiming for faster training and better performance. For example, DETR-SMCA[3] proposed a spatial-aware cross-attention to accelerate the convergence of training. Anchor DETR[20], provides anchor points as prior information for object queries. DAB-DETR is one of the representation works of query design, it reveals that the slow convergence of DETR is due to the multi-pattern nature of object queries. To address this, the proposed DAB-DETR (Dynamic Anchor Box DETR) introduces a novel approach by using 4D anchor boxes (x, y, w, h) as prior information to force queries to learn a detection pattern for a single object, then updated them layer-by-layer. This method provides better spatial priors, considering both the position and size of each box, deepening the understanding of queries in DETR. It not only accelerates DETR's training but also adapts the positional prior to match objects of varying scales, noting a significant advancement over previous works. DN-DETR[11] is an enhanced version of DAB-DETR. It accelerates the training phase via providing model noised boxes to help object query better capture the position of objects. It is one of the best DETR and the denoising process does not change the structure of DAB-DETR, so we take DN-DETR as our baseline model.

Multi-scale representations in DETR Deformable DETR[12] is the first work introducing the multi-scale structure to the DETR family. It modifies the standard DETR model by introducing deformable attention modules. These modules enable the model to focus attention on a set of key sampling points around the reference points, rather than the entire image. Such selective attention mechanism is particularly beneficial for detecting small objects. More importantly, it reduces the computation cost of multi-head attention, so that the deformable attention can extended to multi-scale deformable attention. As a result, DETR can achieve multi-scale based object detection, which is a huge enhancement over single-scale-based model.

Recently, RT-DETR[16] achieves state-of-the-art performance on real-time object detection. It beats yolo family which have dominated the real time detection for years, demonstrating the great potential of DETR. In this paper, author did very elaborate experiments on multi-scale features, finding out features fusion appears in the transformer encoder of Deformable DETR is not optimal. And the FPN like structure[4, 9, 12] tend to have better accuracy with lower computational cost. Consequently, they proposed the attention-based Intra-scale Feature Interaction (AIFI) module for high level features and the CNN-based Cross-scale Feature-fusion Module (CCFM) for other features which achieves an optimal multi-scale structure for DETR. RT-DETR achieves state-of-the-art benchmarks and sets a new standard for multi-scale DETR, nevertheless, it remain the same decoder design as Deformable DETR. Although it is computa-

tionally efficient, but the deformable attention in the transformer decoder can only catch local information around the sample point, impairing model's ability to capture long term dependency. Our proposed method resize the feature maps and rearrange the feature maps, thus we can keep the dense global attention without losing efficiency.

3 Proposed Method

3.1 Overview

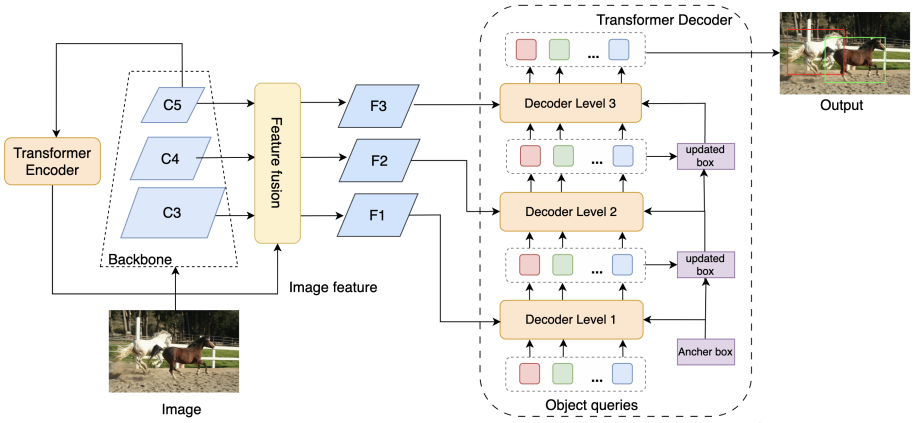


Figure 3: Overall architecture of our proposed network. The image feature from transformer encoder is passing to the bottom of feature fusion module, which is fused with C3. And then the generated feature F1, F2, F3 are feeding into different layers of transformer decoder.

The overall architecture of the proposed work is shown in Fig.3. It contains four parts: a backbone network, a feature fusion module, a transformer encoder and a transformer decoder. We feed multi-scale features $C3$, $C4$ and $C5$ from the backbone and features generated by the transformer encoder into the feature fusion module. Then we arrange the generated features into different transformer decoder layers. The backbone and the transformer encoder are the same as vanilla DETR.

3.2 Feature Fusion Module

Multi-scale features are the ideal candidates to feed into the transformer decoder to solve the inefficient issue. The multi-scale structure has been widely regarded as a valuable component for vision tasks[8] since 1960. For modern deep neural networks, there is a deep representation of images due to the substantial depth of the network. Feature Pyramid Network (FPN)[12] defined the **network stage** as layers that produce output maps of the same size and denoted the feature

maps of the last three stages of ResNet50 as $C3$, $C4$ and $C5$; we follow the same terms in our work for clarity.

To achieve simplicity and efficiency, our feature fusion module borrows basic ideas from the FPN, which fuses features via element-wise addition and lateral connections. However, it has substantial differences compared to FPN: 1. Instead of interpolating high-level feature maps to a larger spatial size and generating feature maps in various scales, we employ a straightforward maxpooling operation to downsample the low-level feature map, resizing the feature maps into the same scale to make it feasible to perform element-wise additions across different feature maps. 2. In contrast to the traditional FPN, our work fuses features in the opposite order to generate features, as illustrated in Fig.4. From our analysis experiments, such structure can achieve optimal performance. Specifically, we use the image feature generated by the transformer encoder, which is the semantically strongest representation, as the base feature. Note that this image feature has been flattened into 3-dimension which is able to be processed by transformer, but we still measure its spatial shape as $\frac{H}{16}$ and $\frac{W}{16}$ for convenience, in which H and W are original size of image. We add the base image feature with reshaped $C3$ (denoted as $C3'$) to generate an enhanced feature map $F1$. The reshaping operation consists of a maxpooling operation to downsample spatial resolution of $C3$ ($\frac{H}{4}$, $\frac{W}{4}$) by a factor of 4 and a linear projection layer to align the number of channels in C3 with our base map. Unlike the FPN, we do not perform another 3×3 convolution on the enhanced feature because we intend to remain the features from transformer unchanged. $F2$ and $F3$ are generated following the same manner. We concludes the whole process as follows:

$$F_1 = x + \text{Conv}_{1\times1}(\text{maxpooling}(c_3)), \tag{1}$$

$$F_2 = F_1 + \text{Conv}_{1\times1}(\text{maxpooling}(c_4)), \tag{2}$$

$$F_3 = F_2 + \text{Conv}_{1\times1}(c_5), \tag{3}$$

where x represents the image features generated by the transformer encoder. Note that We do not need to reshape $C5$ because it has the same size as the image feature. All the feature maps from the backbone will be flattened during the fusion process. In this way, we create features that contain different semantic information which will be used in the different decoder layers.

3.3 Proposed Transformer Decoder

For the naive detection transformer decoder, there are six identical decoder layers. In this work, we evenly arrange six decoder layers into three levels, feeding $F1$, $F2$ and $F3$ generated by our customized feature fusion module into three different decoder levels. Feature $F1$ is enhanced by shallow-stage feature $C3$, and we feed it into first level of the decoder (layer 1 and layer 2). Then

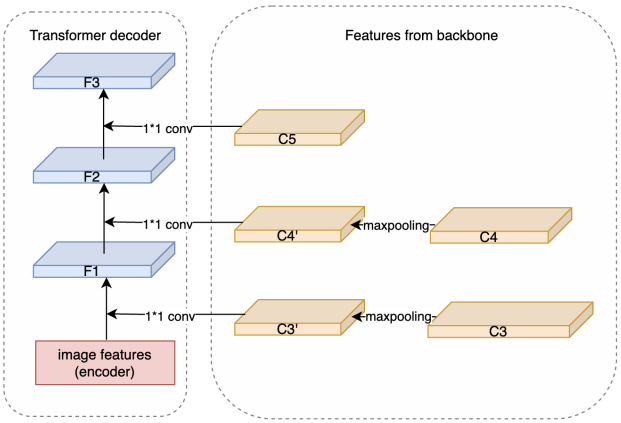


Figure 4: Multi-scale fusion module. Image features are generated by transformer encoder. Note that we downsample C4 and C3 into C4' and C3' respectively, which have the same size as image features.

we feed F2 and $F3$ successively into the next two decoder levels as they contain semantically stronger information. Unlike any previous works that concatenate multi-scale features and feed them into decoder layer simultaneously[12, 16, 22, 24], we set them in cascade in decoder. In this way, we can add abundant semantic information into different decoder levels, fulfill the role of dynamic anchor box, and solve the inefficiency issue we discussed before. The query design is the same as the original DAB-DETR: it provides four-dimensional prior information to object queries and then updates it layer by layer. Moreover, it also modulates the positional attention maps by dividing the relative anchor width and height from its x and y part separately to match with objects of different scales:

$$\begin{aligned} &\text{ModulateAttn}((x,y),(x_{\text{ref}},y_{\text{ref}})) = \\ &\left(PE(x)\cdot PE(x_{\text{ref}})\frac{w_{q,\text{ref}}}{w_q} + PE(y)\cdot PE(y_{\text{ref}})\frac{h_{q,\text{ref}}}{h_q}\right)/\sqrt{D}, \end{aligned} \tag{4}$$

where $(x,\ y)$ and $(x_{ref},\ y_{ref})$ are coordinate of two points, w_q and h_q are the width and height of the anchor A_q, and D is the scaling factor. PE denotes position encoding which is discribed in [2]. $w_{q,ref}$ and $h_{q,ref}$ are the reference width and height that are determined by a shared MLP at different decoder layers:

$$w_{q,ref}, h_{q,ref} = \sigma(\text{MLP}(C_q)), \tag{5}$$

where C_q is content embedding (object query) and activation function σ is chosen as Relu in this paper. Because in our proposed structure object queries interact with different feature maps, we implement a layer-aware ModulateAttn that learns $w_{q,ref}$ and $h_{q,ref}$ by independent MLP in different levels rather than a shared one. However, we only observe a slightly increase on accuracy compared to the baseline which uses only one MLP, demonstrating that the behaviors of object queries are highly related at different levels even though the feature maps are different from each other. Thus in our work we choose to use a shared MLP across different levels for efficiency.

4 Experiments

4.1 Dataset and model parameters

Following the common practice with other state-of-the-art DETR models, we trained our model on the COCO train2017 dataset and evaluate it on the COCO val2017. We use the Mini COCO for ablation experiments. Model parameters settings are almost the same as DN-DAB-DETR for consistency as it serves as our baseline model.

4.2 Training settings and model comparison

The experiments results are shown in Table.1. Our training process aligns with most of the DETR settings for fair comparison. We use AdamW as our optimizer with weight decay 10^{-4} and set learning rate as 1×10^{-4} to train the model for 50 epochs, which drops by 0.1 after 40 epochs. The number of object query is selected as 300. The proposed model is trained on a single node with 2 Nvidia GTX Titan X GPUs, and the batch size for each GPU is 2 for memory consideration thus the total batchsize is only 4, which has a huge gap with DN-DETR from original paper. For a fair comparison, we replicated DN-DAB-DETR with corresponding training settings and hyperparameters. We have also provided our source code at `https://github.com/EddyPianist/detr_mul` for reproduce purpose. In Table.1, we use $*$ to indicate the replicated experiments. We also compare our model with other single scale state-of-the-art DETR which trained on better hardware. Also note that all the backbone of DETRs listed for comparison are ResNet50. Comparing to DN-DETR, the detection accuracy of our proposed model increases by around 3.1% (from 41.9% to 45.0%) with little extra computational cost. Moreover, we also observe a faster convergence in the early stage. Our proposed model can achieve similar accuracy with only 30 epochs compared to the our baseline model. Furthermore, our model outperforms other state-of-the-art DETR such as Anchor DETR, Conditional DETR and DAB-DETR, which require more advanced hardware such as Nvidia A100 or V100.

To comparing with deformable DETR family, we also trained a MLD-DETR-DC5 model, note that although DC5 model use 2 times more resolution fea-

tures than baseline MLD-DETR, they are much smaller than features from DN-deformabel-DETR. We use the same training settings as previous model. Without any other boosts such as top-K query selection[16] or collaborative hybrid assignments[25], our proposed method achieves a better overall precision with less parameters compared to deformable models. Moreoever, our model outperform deformable models in detecting large object. This is because, in the decoder of deformable DETR and other multi-scale DETR models, dense global attention is replaced by sparse local attention, which harms the model's ability to capture global information. In contrast, our proposed method retains dense global attention. Thus, compared to multi-scale deformable DETR, MLD-DETR also demonstrates its advantages.

Model	epochs	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	params
DN-DAB-DETR*	50	41.9	61.9	44.7	21.5	45.6	60.0	44 M
MLD-DETR (ours)	50	45.0	65.0	48.2	24.4	49.2	63.3	44 M
MLD-DETR(ours)	30	41.9	62.1	44.6	21.8	45.2	60.6	44 M
DETR	500	42.0	62.4	44.2	20.5	45.8	61.1	41 M
Anchor-DETR	50	42.1	63.1	44.9	22.3	46.2	60.0	39 M
DAB-DETR	50	42.2	63.1	44.7	21.5	45.7	60.3	44 M
Conditional DETR	50	40.9	61.8	44.3	20.8	44.6	59.2	44 M
Deformable-DETR	150	43.8	62.6	47.7	26.4	47.1	58.1	40 M
SCMA-DETR	50	43.7	63.6	47.2	24.2	47.0	60.4	40 M
DAB-Deformable-DETR	50	46.8	66.0	50.4	29.1	49.8	62.3	47 M
DN-Deformable-DETR*	50	46.7	65.4	50.7	29.0	48.4	62.5	48 M
MLD-DETR-DC5(ours)	50	47.0	66.8	50.4	27.9	50.7	64.7	44 M

Table 1: Performance comparison of different DETR. The $*$ indicates our replicated experiments on our local hardware for a fair comparison.

4.3 Detection accuracy of different decoder layers

In this subsection, we conduct ablation experiments of different decoder layers to demonstrate that the proposed method successfully solve the inefficiency issue in the DETR transformer decoder. The experiment itself is straight-forward: In DETR, the final prediction is computed by a 3 layers Feed Forward Network(FFN). Additionally, the prediction loss is computed at each transformer decoder layer with the shared FFN for better performance. Thus we evaluate the results using output from the shared FFN at different layers on COCO eval to observe the detection accuracy of different layers. The experiment results are shown in Fig.5. It demonstrates that our proposed method already has advantages over baseline model from the early stage. And the performance gap increase significantly at decoder layer 3 when we introduce F2 to the fused feature. We infer that feature $C4$ plays an important role as it bridges the semantic gap between $C5$ and image features from the transformer encoder, so that boosting the detection accuracy dramatically at the decoder layer 3. Moreover, for

layer 5 and 6, we observe modest accuracy improvement at baseline(0.9%). On the contrary, the proposed MLD-DETR shows a decent improve in the last two layers (1.4%). Thus, we can conclude our multi-level decoder design resolves the inefficiency issues present in other DETR transformer decoders.

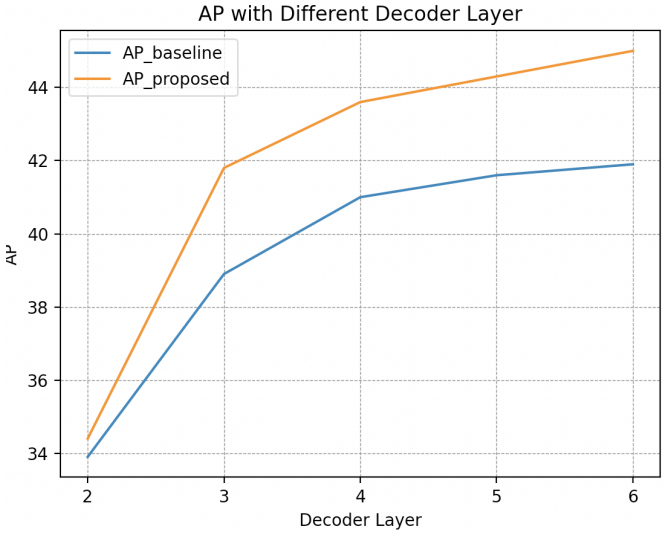


Figure 5: Comparison on MLD-DETR's detection accruacy and DN-DETR's detection accuracy with different decoder layers

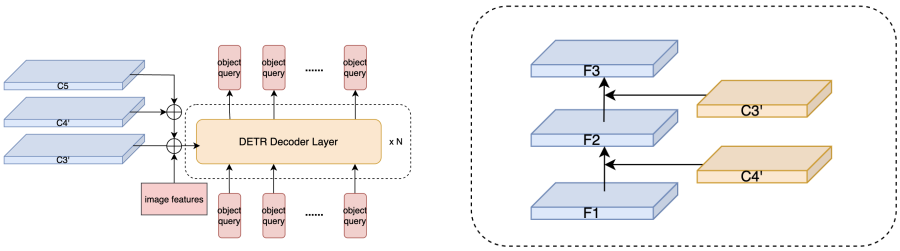
4.4 Ablation Study on Mini COCO

Overview Although we have observed that proposed model demonstrates significant improvement over the baseline model and improve the efficiency of the DETR transformer decoder, there are several questions remained to be addressed: 1. The improvements may not come solely from multi-scale features and multi-level design; the additional information provided by the backbone itself could be the reason for the accuracy improvement. 2. The way we fuse feature maps is not aligned with the principle of the Feature Pyramid Network (FPN), which ensures the addition of features with minimal semantic gaps. It is critical to verify if the naive FPN structure could work in our proposed decoder structure. Therefore, it is necessary to conduct several analysis experiments to address these questions and provide a deeper insight into the proposed model.

We perform different experiments on the Mini COCO dataset and validate them on the original COCO validation set. Mini COCO is a subset of the COCO2017 training dataset which contains only 25K images. It is 10 times smaller than original MS COCO dataset[13]. Elaborate experiments are conducted in [19], confirming that detection models trained on Mini COCO had similar trends

compared to those trained on original dataset, so in this work we use it to efficiently conduct ablation experiments.

Integrity of multi-level structure Firstly, to verify that the improvements truly come from multi-level design rather than simply from additional information from the backbone feature maps, we design a structure that mixes the multi-scale at once, which denoted in Fig.6 (a). We use the same maxpooling operation to reshape features, then we add them simultaneously to generate one feature map, which is fed to all encoder layers, to see if it can boost our detection accuracy. The result is shown in Table.2, we do not observe any improvement; instead, the accuracy dropped to 31.2% compared to baseline DN-DETR (33.4%, trained with minicoco dataset). Thus simply stacking the information from the backbone is harmful for object queries to detect objects.



(a) DETR decoder with mixed features (b) FPN style feature fusion

Figure 6: Comparision between our proposed multi-scale method and traditional multi-scale method in DETR

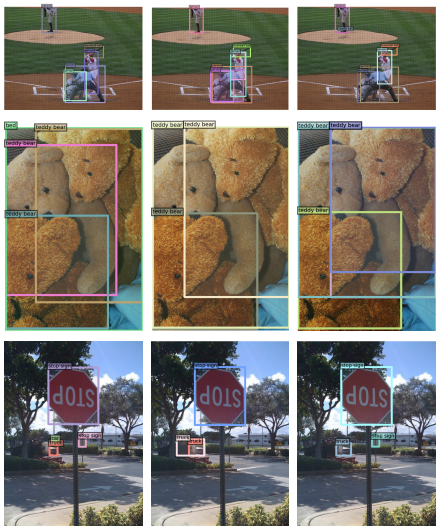
Effect of feature fusion module Then, to testify our feautre fusion design, we set a comparison experiment that strictly follows the pipeline of the FPN(as shown in Fig.6(b). Instead of $C3$, we fuse high-level feature $C5$ with features generated by the encoder. Then the generated feature fuses with downsampled $C4$ and $C3$ sequentially, to ensure we only adding features with similar semantic information. Note that we do not change the size of feature maps but only the order in which we arrange them. Although this structure contains multi-level features and seems to have minimum changes over our proposed one, it does not show any improvements. In fact, compared to the baseline DN-DETR, the detection accuracy is almost the same. Consequently, based on our analysis experiments, we can conclude that our feature fusion module is effective and feeding features with different semantic information into different decoder level is critical in our design. Moreover, it also distinct our work from works that involved FPN-like architecture such as RT-DETR and CF-DETR.

Other Ablation studies Feature map $C5$ is fused at the last decoder level to enhance semantic information of previous features. It takes most of our computational resources in proposed feature fusion models due to the large number of channels of $C5$. Without it, the total number of parameters drops from 44.3M to

Model	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	params
DN-DAB-DETR	33.4	53.2	34.7	13.9	35.8	51.1	43.7M
MLD-DETR	34.6	54.1	36.1	15.6	37.3	51.3	44.3M
Decoder w/ mixed features	32.2	51.5	33.9	13.3	35.1	49.9	44.3M
Decoder w/ FPN	32.7	52.6	34.2	47.7	35.6	15.2	44.3M
MLD-DETR(C5 eliminated)	34.2	53.8	36.0	15.3	37.2	50.3	43.8M
Scaled-MLD-DETR	32.3	52.1	34.2	15.2	35.5	49.0	49.0M

Table 2: Ablation studies on Mini COCO.

43.8M, which is almost the same as our baseline model (43.7M for DN-DETR). Moreover, the feature of $C5$ is semantically strong which might introduce redundant information because we already have the semantically strong base features. To assess the impact, we removed the feature fusion from the last level and observed how it affects accuracy. From the results of ablation experiments, feature fusion in the last two decoder layers is somehow critical: the detection accuracy will drop from 34.6% to 34.2%, counting almost 1/3 of our accuracy improvement. Additionally, we can observe that the detection accuracy for large objects dropped significantly, which implies the semantic information from $C5$ is crucial for large object detection.



(a) (b) (c)

Figure 7: Visualization on some instances of COCO2017 val.(a)ground truth, (b)DN-DETR, (c)MLD-DETR

4.5 Comparison on Visual Results

In this subsection, we present visual results of the baseline model, our proposed model, and the ground truth for reference. As shown in Fig.7, our proposed method demonstrates the ability to detect challenging objects. For example, in the first row, there are multiple people gathered together in different postures, which is difficult for DN-DETR to distinguish, but our model can correctly detect them. Additionally, DN-DETR fails to accurately detect three teddy bears due to their similar texture, whereas MLD-DETR succeeds. Remarkably, as shown in the last row, the proposed model can detect a stop sign from its back, which could be challenging even for a human. We attribute this to the introduction of low-level semantic information into the decoder, so that the model can identify the stop sign by its shape. Therefore, these visualized instances demonstrate that incorporating various semantic information improves the model's ability to capture difficult objects.

4.6 Limitations

Our proposed method solves the inefficiency in the transformer decoder by introducing abundant semantic information into it. A natural question that arises is whether adding more decoder layers could further improve detection accuracy. To answer this question, we did another ablation experiment expanding the number of decoder layers from six to nine: we added one layer to each different decoder level (thus we have three F1 layers, three F2 layers, and three F3 layers). However, the detection transformer can not benefit from more decoder layers, the accuracy of the 9-layer decoder we have observed was almost the same as the one with 6 layers. We admit that the result is not consistent with our intention, but we still present this result in order to inspire further exploration into this topic.

Model	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	params
MLD-DETR	34.6	54.1	36.1	15.6	37.3	51.3	44.3M
MLD-DETR w/ 9 layers	34.5	53.7	35.9	15.5	37.1	51.0	48.0M

Table 3: Experiments with more decoder layers and scaled feature maps. Note that experiments shown in this table also conducted on Mini COCO

5 Conclusion

In this paper, we introduce Detection Transformer with Multi-Level Decoder. It fuses multi-scale features and arrange them vertically at different layers of decoder, improving the efficiency at the transformer decoder side. Furthermore, it provides a new perspective to the application of multi-scale features in DETR model. The proposed method improve the detection accuracy by 1.4% than its

baseline DN-DAB-DETR on MS COCO dataset with almost no extra computational cost. Nevertheless, we found out our model does not work as expectation with scaled feature maps or more decoder layers. We anticipate that the DETR research community will explore more on related topics.

Data Availability

The training script and checkpoints of this study are openly available at `https://github.com/EddyPianist/detr_mul`.

References

[1] Xipeng Cao, Peng Yuan, Bailan Feng, and Kun Niu. Cf-detr: Coarse-to-fine transformers for end-to-end object detection. In *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, 2022.

[2] 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 (ECCV)*, 2020.

[3] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. *arXiv preprint arXiv:2101.07448*, 2021.

[4] Golnaz Ghiasi, Tsung-Yi Lin, Ruoming Pang, and Quoc V. Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *CVPR*, 2019.

[5] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] DH Hubel and TN Wiesel. Receptive fields of optic nerve fibres in the spider monkey. *J Physiol*, 154(3):572–580, 1960.

[9] Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. Parallel feature pyramid network for object detection. In *ECCV*, 2018.

[10] Feng Li, Ailing Zeng, Shilong Liu, Hao Zhang, Hongyang Li, Lei Zhang,

and Lionel M Ni. Lite detr: An interleaved multi-scale encoder for efficient detr. *arXiv preprint arXiv:2303.07335*, 2023.

[11] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dndetr: Accelerate detr training by introducing query denoising. *arXiv preprint arXiv:2203.01305*, 2022.

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[13] 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 *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[14] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

[15] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[16] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu. Detrs beat yolos on real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. *arXiv preprint arXiv:2108.06152*, 2021.

[18] S. Ren, Kaiming He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NIPS)*, 2015.

[19] N. Samet, S. Hicsonmez, and E. Akbas. Houghnet: Integrating near and long-range evidence for bottom-up object detection. In *ECCV*, 2020.

[20] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2567–2575, 2022.

[21] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: Improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.

[22] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes

for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2022.

[23] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In *AAAI*, 2019.

[24] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

[25] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the International Conference on Computer Vision (ICCV)*. Computer Vision Foundation, 2024.