PAPER

DFAM-DETR: Deformable Feature Based Attention Mechanism DETR on Slender Object Detection

Feng WEN[†], **Mei WANG**[†], ***and*** **Xiaojie HU**[†a)], ***Nonmembers***

SUMMARY Object detection is one of the most important aspects of computer vision, and the use of CNNs for object detection has yielded substantial results in a variety of fields. However, due to the fixed sampling in standard convolution layers, it restricts receptive fields to fixed locations and limits CNNs in geometric transformations. This leads to poor performance of CNNs for slender object detection. In order to achieve better slender object detection accuracy and efficiency, this proposed detector DFAM-DETR not only can adjust the sampling points adaptively, but also enhance the ability to focus on slender object features and extract essential information from global to local on the image through an attention mechanism. This study uses slender objects images from MS-COCO dataset. The experimental results show that DFAM-DETR achieves excellent detection performance on slender objects compared to CNN and transformer-based detectors.

key words: *slender object detection, Deformable DETR, DFAM, deformable convolution, attention mechanism*

1. Introduction

Object detection has made significant development in recent years as a crucial study subject, with the increasing application of deep learning on computer vision [1]–[4]. For each object of interest in an image, object detection needs the algorithm to predict a bounding box with a category label. One-stage detectors and two-stage detectors are the two primary types of object detectors. For instance, the YOLO series [5]–[8], SSD [9], DSSD [10], Retina-Net [11], Efficient-Det [12], FCOS [13], and Corner-Net [14] are one-stage detectors with the benefit of detection speed. The R-CNN [15], Cascade R-CNN [16], Fast R-CNN [17], and Faster R-CNN [18] are two-stage detectors with the advantage of high detection accuracy.

The demand for object detection, such as small object detection [19] and dense object detection [20], [21], is growing as the field of computer vision develops. Certain detecting effects have been obtained, and some novel approaches and solutions have been offered. Despite the fact that the most of the issues have been resolved, there is still a significant difference between slender object detection and regular object detection. For slender object detection, the previously described one-stage and two-stage detectors, such as Faster R-CNN [18], RepPoint [22], and FCOS [13], were used. The test included only slender object images from the MS-COCO dataset [23], such as knives, forks, skis and snowboards. The highest detection accuracy of AP reached 20.7 percent. This is because of the standard convolution only samples the input feature map at fixed locations and it cannot automatically adjust the sampling points to fit the features of slender objects.

Since Transformer's self-attention layers are global instead of locality two-dimensional neighborhood structure, it has much less image-specific inductive bias than CNNs [24]. Research such as Detection Transformer (DETR) [25] started to applying transformer for object detection. Results show that the attention mechanism in the transformer has strong modeling capability for relation. The main target area is obtained by scanning the global image. It effectively concentrates on the image's slender object for improved output quality. Furthermore, the data dimension is reduced which can lower the computational load of high-dimension data input.

The downside of DETR is that using the transformer attention mechanism to obtain sampling points is still time demanding. Deformable DETR [26] successfully integrates transformer and deformable convolution [27] with sparse spatial sampling positions to solve the problem of slow convergence speed and high complexity of DETR. The deformable attention module in Deformable DETR only obtain key sampling points around a reference. Convergence and feature spatial resolution issues can be reduced by allocating a fixed number of important points to each query. It can provide efficient and better detecting performance with fewer and more precise sampling points on the slender objects.

Furthermore, while Deformable DETR merely adds a deformable attention module to the transformer, backbone network feature extraction is still insufficient for detecting slender objects. When CNN is used to extract features in the backbone network, it has difficulty adapting to the shape of slender objects. Hence, we propose Deformable Feature based Attention Mechanism DETR (DFAM-DETR) detector for slender object detection. This detector is based on Deformable DETR, and Deformable Feature based Attention Mechanism (DFAM) is designed to sample slender object features and increase the ability of feature extraction by deformable convolution and attention mechanism. Deformable convolution can adjust the position of sample points in the image adaptively. It assures that the sample points are localized in the image's region of interest to avoid background influence. For instance, as illustrated in

Manuscript received June 22, 2022.
Manuscript revised November 9, 2022.
Manuscript publicized December 9, 2022.
[†]The authors are with School of Information Science and Engineering, Shenyang Ligong University, Liaoning, China.
a) E-mail: xiaojie.hu@sylu.edu.cn
DOI: 10.1587/transinf.2022EDP7111

Fig. 1 Sampling points for slender objects. (a) Standard convolution sampling points; (b) DETR sampling points.

Fig. 1 (b), using deformable convolution with DFAM may cluster the sampling points more accurately on slender object than the standard convolution in Fig. 1 (a).

Apart from the deformable convolution, the attention mechanism is also applied with DFAM for slender object detection. The attention mechanism may learn which information to emphasize or suppress based on the dimensions of the channel and space. Hence, it increases the effectiveness of recognizing slender objects by focusing on important features and suppressing those that aren't. In summary, the proposed DFAM-DETR detector is modified based on Deformable DETR specifically for slender object detection. The DFAM is designed for capturing the specific features of slender objects. As a result, DFAM-DETR detector greatly improves slender object detection accuracy and efficiency comparing to Deformable DETR.

2. Related Work

Object detection is categorized into one-stage detectors and two-stage detectors [1]–[4]. One-stage detectors do not require the region proposal stage and may generate the probability of an object's category and location directly. For instance, the YOLO series [5]–[8], SSD [9], DSSD [10], Retina-Net [11], Efficient-Det [12], FCOS [13], and CornerNet [14] are typical one-stage detectors with the benefit of detection speed. The two-stage detectors must first create region proposals, then perform object classification and localization for region proposals. For instance, the R-CNN [15], Cascade R-CNN [16], Fast R-CNN [17], and Faster R-CNN [18] detectors have the benefit of high detection accuracy.

Despite the fact that object detection using convolution has gained high accuracy, detection performance on slender objects remains poor. The convolutional approach has difficulty capturing features of slender objects. Popular object detectors like Faster R-CNN [18], FCOS [13], RepPoints [22] adopt standard convolution. Furthermore, improved detectors [28] based on FCOS and RepPoints that specifically developed for slender object detection still shows weak detection accuracy.

The self-attention mechanisms of transformer can scan through each element of a sequence and update it by aggregating information from the whole sequence [25], [26], [29]–[32]. DETR is a transformer-based object detector. It combines the bipartite matching loss and transformers with powerful relationship modeling ability [25]. However, DETR requires more epochs to achieve convergence comparing to popular detectors [26]. Due to the complexity of high-resolution feature map, the performance of DETR in detecting small objects is relatively poor [26]. Deformable DETR is an effective and efficient detector for dealing with sparse spatial locations, which compensates the lack of the element relation modeling capability for DETR [26], [27]. The deformable attention module of Deformable DETR only focuses on a small group of key sampling points around the reference point without considering the spatial size of the feature map [26]. By allocating only a few fixed numbers of keywords to each query, the problems of convergence and spatial resolution of elements can be alleviated [26].

However, Deformable DETR only introduces deformable attention module into transformer. For slender objects, feature extraction of backbone network still adopts convolution, which does not provide sufficient solution. Hence, we propose DFAM based on deformable convolution feature and attention mechanism for effective slender object detection. Unlike Deformable DETR, DFAM use adaptive sampling points of deformable convolution and attention mechanism to aggregate the whole input sequence information in backbone network to accurately identify slender objects and obtain better detection accuracy for slender objects.

3. Method

3.1 Architecture

DFAM-DETR is based on Deformable DETR, which is comprised of three parts, the ResNet [1] as backbone, the transformer with encoder-decoder, and the Feed Forward Network (FFN). As shown in Fig. 2, our improvement is mainly in backbone network. The Deformable Feature based Attention Mechanism (DFAM) is designed in the backbone network based on ResNet to extract slender object features. Transformer takes full advantage of its powerful modeling capabilities and sampling capability of deformable attention module to improve the accuracy of slender object detection. FFN is used to predict the output categories and position of objects in the picture.

3.2 Feature Extraction of Backbone

The backbone is anticipated to fully mine the meaningful semantic information of the image as the model's core feature extraction function. Convolution in the ResNet backbone is challenging to adapt to the unique shape of slender objects. We propose the DFAM to enhance the ability of feature extraction by improving one layer (C5) in ResNet, see Fig. 2.

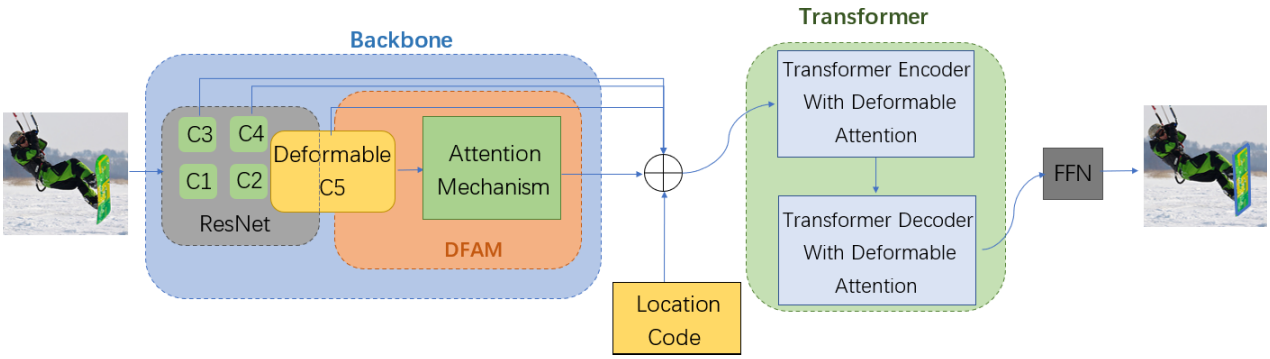


Fig. 2 Overall structure of DFAM-DETR. Note: C1, C2, C3 and C4 are original from ResNet.

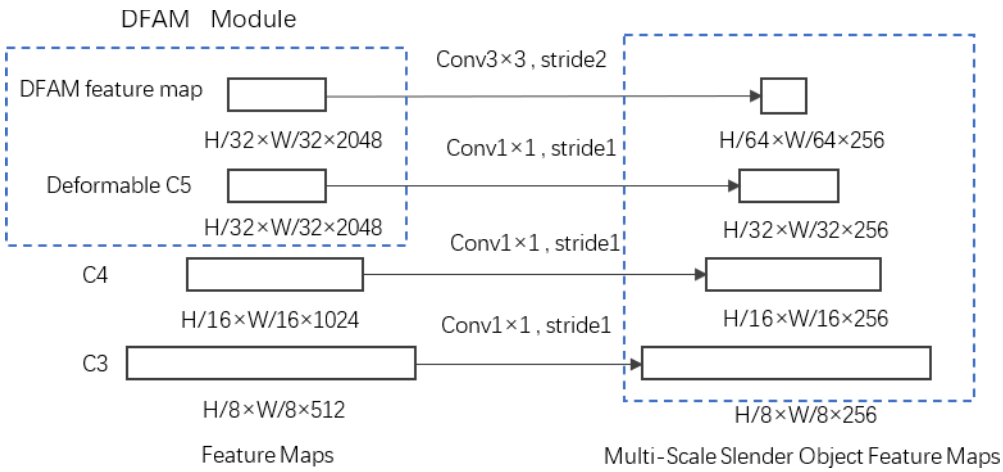


Fig. 3 Constructing multi-scale slender object feature maps

First, ResNet generates feature maps for stages 3, 4, and 5 as C3, C4, C5. To build the deformable C5 feature map, we replace the convolution in the last stage of ResNet with deformable convolution, which is then fed into the attention mechanism to generate the DFAM feature map. Second, in order to capture the different scales of slender objects, the C3, C4, deformable C5 and DFAM feature map are adapted to generate multi-scale slender object feature maps. The C3, C4, and deformable C5 feature maps are convolved with the 1×1 stride 1 to get the first three feature layers. Then the last layer feature map is obtained via a 3×3 stride 2 convolution on the DFAM feature map, see Fig. 3. Therefore, the multi-scale slender object feature maps are captured from the backbone. Lastly, the multi-scale slender object feature maps are input to transformer to enhance the ability of semantic and geometric information representation.

3.3 The Proposed Deformable Feature Based Attention Mechanism

As shown in Fig. 4, DFAM combines the ability of deformable convolution's adaptive sampling points with the capacity of focusing critical features of the attention mechanism to adjust to the features of slender objects and increase feature extraction ability.

The geometric structure and sample points of the standard convolution kernel are fixed in the convolutional neural network, and the generalization capacity is limited, thus the geometric modification has inherent restrictions. Because of this constraint, a model can only get feature information from a fixed area, making it impossible to adjust to the feature of slender objects and limiting their capacity to extract features. As a result, we propose using deformable convolution instead of standard convolution in the backbone to extract features of slender objects. The comparison between standard convolution sampling points and deformable convolution sampling points in slender object is showing in Fig. 5. Comparing to standard convolution, the deformable convolution includes a learnable offset at each sampled position in the feature map so that the deformable convolution can better adapt to the features of slender objects. Furthermore, while deformable convolution does not considerably increase the model's parameters and FLOPS, too many deformable convolution layers would dramatically increase the infer time in reality. As a result, in order to balance efficiency and accuracy, we propose replacing 3×3 convolution layers in the last stage C5 with 3×3 deformable convolution layers.

Deformable convolution can adjust spatial samples with extra offsets and learn the offsets of target tasks

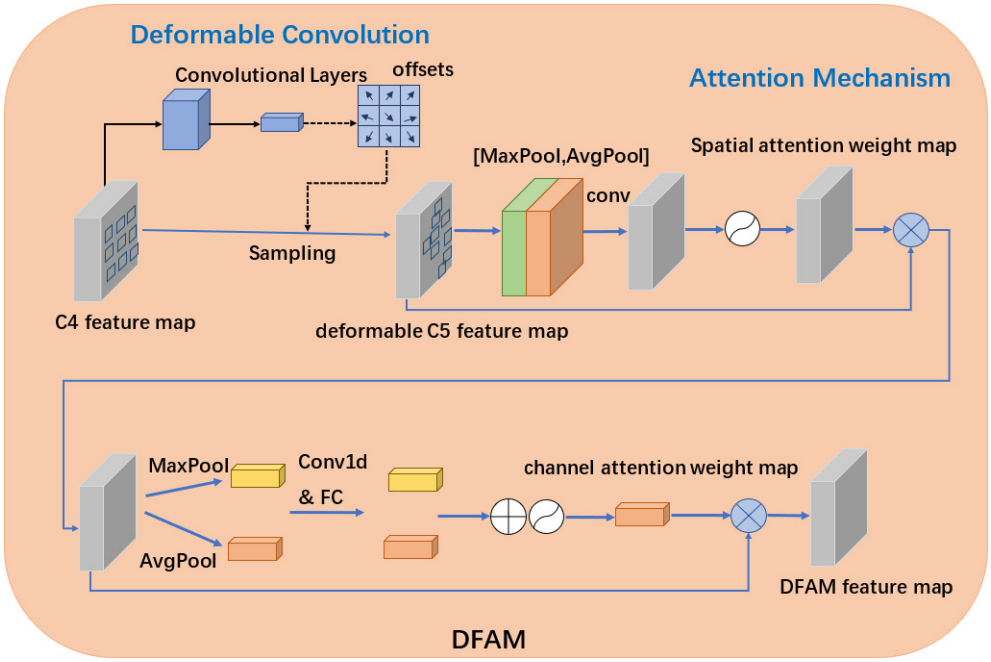


Fig. 4 DFAM framework.

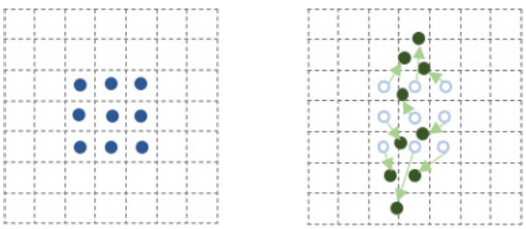


Fig. 5 Comparison of 3×3 standard and deformable convolution sampling points. (a) Standard convolution sampling points (dark blue points); (b) deformable convolution sampling points (dark green points).

without additional supervision. As shown in Fig. 4, the two-dimensional offset can be calculated by another parallel standard convolution unit, and can also be learned end-to-end by gradient back propagation to generate new sampling positions in the feature map [33]. For a sampling field R with the size of $(N \times N)$, $R = \{(0,0),(0,1),\ldots,(N-1,N-1)\}$, and an input image data x, for each location p_0 on the output feature map y_{deform}, the formula of deformable convolution is as follow [34],

$$y_{\text{deform}}(p_0) = \sum_{p_n \in R} w(p_n) * x(p_0 + p_n + \Delta p_n) \tag{1}$$

where p_n is all the locations in the sampling field, and Δp_n is the offset position of each sampling point, and $w(p_n)$ is the corresponding weight [34],

$$\Delta p_n = \left(x_{offset}, y_{offset}\right) \mid (0,0),\ldots,(N-1,N-1) \tag{2}$$

where $(x_{\text{offset}}, y_{\text{offset}})$ represents the offset of x coordinate and y coordinate of a certain position respectively.

Since the offset Δp_n is typical fraction, and discrete image data cannot use non-integer coordinates, bilinear interpolation is adopted in Eq. (1). The intuitive effect is that the deformable convolution can adjust the position of sampling points according to the feature of the slender object.

It is crucial to show the channel content and space location of the slender object in the image instead of the background area. The attention mechanism concentrates on the features of slender objects in the image and ignores those that are irrelevant. Hence, we propose a deformable convolution-based attention mechanism for better focusing on the features of slender objects. The proposed attention mechanism consists of two dimensions of channel and space to better extract the features of slender objects, see Fig. 4.

Spatial attention focuses on activated spatial information in the feature map [35], which can enhance the valuable local spatial information while suppressing the slender objects' background noise information. The spatial attention feature map is obtained by feeding the deformable C5 feature map through the spatial attention mechanism, see Fig. 4. To generate two one-dimensional feature maps, the average-pooling and max-pooling are used to the deformable C5 feature map first. Second, a two-dimensional feature map is created by concatenating two one-dimensional feature maps. A novel one-dimensional spatial attention weight map is created by convolution with the 7×7 convolution kernel to determine the spatial attention weights of slender objects. To compress the spatial attention weights into a range, the sigmoid function is utilized $(0,1)$. Finally, the spatial attention weight map W is applied to the initial feature map $y_{\text{deform}}(p_0)$ by element-wise multiplication to get the spatial attention feature map. The formulas are defined as follows:

$$f' = \text{Avg}\, P_{sp}\left(y_{\text{deform}}\left(p_0\right)\right) \oplus \text{Max}\, P_{sp}\left(y_{\text{deform}}\left(p_0\right)\right) \quad (3)$$

$$W = \phi\left(f'\right) \quad (4)$$

$$F_{sp} = \sigma(W) \odot \left(y_{\text{deform}}\left(p_0\right)\right) \quad (5)$$

where $AvgP_{\text{sp}}$ and $MaxP_{\text{sp}}$ represent average-pooling and max-pooling, respectively. The symbol ϕ is the convolution layer, and its filter size is 7×7. The symbol $\bigoplus$ is the connection operation on the channel axis. The symbol $\odot$ is the element-wise multiplication on each channel. The symbol σ denotes the sigmoid function.

Channel attention is then used to transmit the spatial attention feature map. The channel attention module sets weights to different dimensions of features so that the ones that contribute the most to the representation of slender object features is highlighted. As illustrated in Fig. 4, the spatial attention feature map is first processed using the average-pooling and max-pooling layers, which can learn statistical information about the input features. Second, the pooling layer's output is processed by a shared network composed of a one-dimensional convolution layer and a fully connected layer, which is then connected by element-by-element addition. Finally, the sigmoid activation function is used. To acquire the final DFAM feature map, the learned one-dimensional channel attention weight map is applied to the spatial attention feature map via element-wise multiplication [35]. The formulas are defined as follows:

$$W = \sigma\left(\varphi_1\left(AvgP_{ch}\left(F_{sp}\right)\right) + \varphi_2\left(\text{Max}\, P_{ch}\left(F_{sp}\right)\right)\right) \quad (6)$$

$$F_{ch} = W \odot F_{sp} \quad (7)$$

where $AvgP_{\text{ch}}$ indicates the operation of average-pooling, while $MaxP_{\text{ch}}$ represents max-pooling. φ_1 and φ_2 denote fully connection layers.

3.4 Transformer Encoder and Decoder

Unlike convolution, which can only obtain local features, transformer employs attention-based mechanisms of encoder and decoder to obtain global to specific features [25], [26], [29]–[32]. We make use of the transformer encoder and decoder. As input, the encoder uses multi-scale slender object feature maps. The transformer layer of encoder conducts multi-head attention to capture the context of global slender object features, which locate the association between different pixels in slender feature map. Object query is introduced in the decoder to narrow down the searching space of objects. Finally, transformer can focus on slender objects in an image. The details of transformer in object detection can be found in Deformable DETR [26].

The sampling points of the encoder and decoder layers are visualized in the transformer part, see Fig. 6. The sampling points represent the weight distribution and accuracy. Sampling points with high weights are highlighted in red, while those with low weights are highlighted in blue. In both the encoder and decoder stages, the results show that DFAM-DETR has better detection accuracy than Deformable DETR. DFAM-DETR sampling points are more adaptable to slender objects and clustered near the slender objects. Moreover, sampling points with higher weights are more focused on the objects.

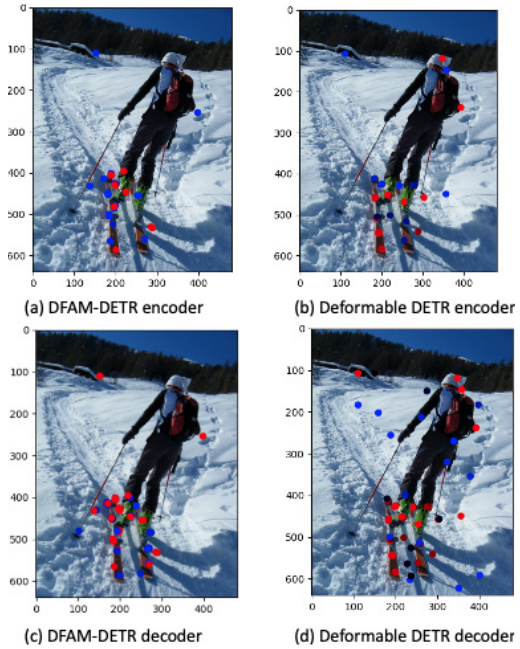


Fig. 6 Visualization of sampling points and attention weights of DFAM-DETR and Deformable DETR.

3.5 Loss Function

In this study, the loss function is consistent with Deformable DETR [26], and the total loss includes classification loss and regression loss. During the model training phase, the Hungarian algorithm [36] is used to match the GT with the model's prediction outcomes. The Hungarian algorithm (bipartite graph matching method) is adopted to determine the optimum arrangement with the least amount of matching loss. The optimal matching result is used to determine the loss function.

4. Experiments and Results

4.1 Dataset

The slender objects dataset used in this study is manually extracted from MS-COCO2017 [23]. It includes slender objects such as toothbrush, snowboard, surfboard, etc. Totally 25,424 training images are used for training, and 1077 images for testing. Data augmentation [19] is performed on the data through random cutting.

4.2 Experiments

4.2.1 Experimental Environment

The experiment was carried out on a Xeon 3104 and an

Table 1 The ablation study of deformable convolution added to different layers.

Modified layer	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
C2, C3, C4, C5	34.4	54.5	36.6	14.6	38.2	**54.8**
C3, C4, C5	34.8	56.6	36.8	14.3	39.1	54.0
C4, C5	34.4	54.8	37.3	12.6	39.4	53.7
C5	**35.4**	**57**	**37.8**	**15.6**	**40**	53.9

Table 2 Effects of different attention mechanisms on the detection accuracy of slender objects.

Modified layer	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
sa	34	54.9	35.4	14.6	39.8	51.4
ca	35	56.4	37.1	14.5	40.7	52.6
sa-ca	**35.4**	**57**	**37.8**	**15.6**	**40**	**53.9**

Table 3 A comparison of the detection accuracy of four detectors for slender objects.

Detector	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN	17.9	35.1	15.7	2.7	17.9	34.5
RepPoints	18.5	34.1	18.1	2.6	18.6	36.1
FCOS	20.7	38.6	20.0	7.5	24.1	31.0
DETR	**28.8**	**49.3**	**28.6**	**8.6**	**31.8**	**51.0**

Table 4 A comparison of the detection accuracy of detectors that based on transformer.

Detector	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR	30.8	52.7	30.3	11.5	33.7	53.0
Deformable DETR	33.4	54.8	34.3	13.0	37.2	**54.5**
DFAM-DETR	**35.4**	**57.0**	**37.8**	**15.6**	**40.0**	53.9

NVIDIA Tesla V100 16GB graphic card, and the environment used Pytorch 1.5.1. The network was trained for 50 epochs using the Adam optimizer [37]. The transformer's initial learning rate was 1×10^{-4}, while the backbone was 2×10^{-5}. The learning rate dropped by 10 times for every 20 training epochs as the number of training epochs increases. The batch size was set to 2, while the weight decay and momentum was set to 0.0001 and 0.9, respectively.

4.2.2 Ablation Study

Deformable convolution of the proposed DFAM-DETR is added to different layers of the backbone network for the ablation experiment, see Table 1. The results show that the accuracy only reaches 34.4% when deformable convolution is added to all layers; 34.4% after adding to C4C5 layers; 34.8% after adding to C3C4C5 layers; and 35.4% after adding to the last layer. Deformable convolution is very effective for detecting slender objects. Adding deformable convolution only to the final layer reaches the best result. The detection effect, however, varies for slender objects at different scales. Especially for large slender object detection, only using high-resolution deformable convolution is insufficient for capturing features.

Ablation experiments were used to assess and compare the effectiveness of channel and spatial attention mechanisms, as shown in Table 2. The results show that channel attention is critical in detecting slender objects. However, using only the channel attention mechanism results in an unsatisfactory detection effect. Therefore, DFAM-DETR includes both channel and spatial attention mechanisms to better extract features and detect slender objects.

4.2.3 Slender Object Detection Results Comparison

In this study, ResNet50 was used as the backbone network. First, the pre-trained Faster R-CNN, RepPoints, FCOS, and DETR detectors were evaluated for slender object detection, see Table 1. Comparing to Faster R-CNN, RepPoint, and FCOS, it revealed that utilizing a detector based on transformer resulted in a significant improvement in the detection accuracy of slender objects. DETR boosted the AP by 10.9 percent comparing to Faster R-CNN, and 14.2 percent for AP_{50}. DETR have significant increase on detection accuracy for both AP_{S}, AP_{M} and AP_{L}. The AP_{S} was increased by 5.9 percent using DETR, AP_{M} was increased by 13.9 percent and AP_{L} was increased by 16.5 percent comparing to Faster R-CNN. The above experiments showed the effectiveness of transformer on slender object detection.

Second, the dataset was trained and evaluated using DETR, Deformable DETR, and DFAM-DETR for comparing the accuracy of slender object detection. ResNet50 was employed for the backbone network. We initialize our backbone networks with the weights pre-trained on ImageNet [38], [39]. Transformer is trained with random initialization. Results showed improvement in detection accuracy on AP, AP_{50}, and AP_{75} with DFAM-DETR, see Table 2. Comparing to DETR, the proposed DFAM-DETR increases slender objects detection accuracy by 4.6 percent on AP, and 4.3 percent on AP_{50}. DFAM-DETR outperformed Deformable DETR by 2 percent increase on AP and 2.2 percent increase on AP_{50}. Furthermore, DFAM-DETR significantly improves detection accuracy for small and medium objects. The results revealed that accuracy was improved by 4.1 and 2.6 percent for small objects, while 6.3 and 2.8 percent increase for medium objects. However, the detection accuracy for large objects was dropped by 0.6 percent comparing to Deformable DETR. This decrease in detection accuracy is primarily caused by the size of the receptive field. Deformable DETR's backbone network utilizes standard convolution layers. The proposed DFAM-DETR, on the other hand, substitutes a deformable convolution layer for the final convolution layer. Deformable convolution has the advantage of being able to adapt its receptive field to the target object. However, due to the size limitation of the receptive field, it is unable to capture the larger slender object in the last layer of the backbone network.

The detection accuracy and convergence curves of Deformable DETR and proposed DFAM-DETR are illustrated in Fig. 7. DFAM-DETR shows higher slender detection accuracy of AP comparing to Deformable DETR. Moreover, DFAM-DETR achieves 2 times less training epochs. As shown in Fig. 8, the training loss of DFAM-DETR is significantly lower than Deformable DETR. Again, it depicts

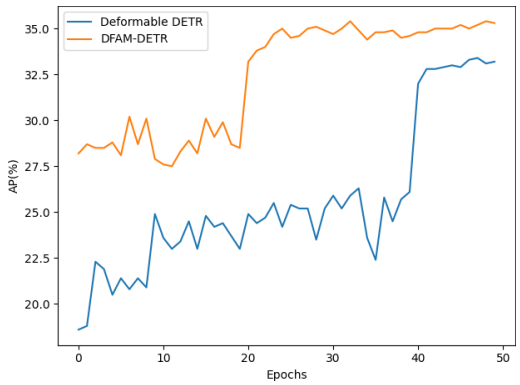


Fig. 7 A comparison of accuracy and epochs between Deformable DETR and DFAM-DETR.

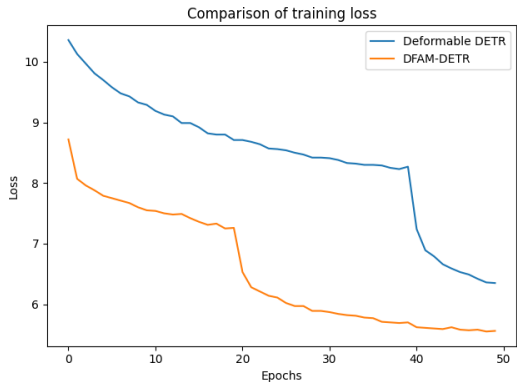


Fig. 8 A comparison of loss and epochs between Deformable DETR and DFAM-DETR.

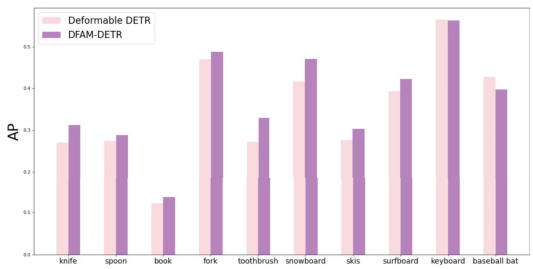


Fig. 9 Comparison of detection accuracy within each category.

that DFAM-DETR has superior convergence speed and it has better performance on slender object detection. The deformable convolution and attention mechanism of DFAM are critical factors in increasing convergence speed. The sampling points are more focused on slender objects comparing to Deformable DETR. This feature extraction mechanism can accelerate overall convergence and loss reduction, making it more suitable for classification and regression.

Figure 9 shows that DFAM-DETR outperforms Deformable DETR in most categories when it comes to detecting slender objects. As a result, the efficacy of DFAM-DETR is confirmed.

The ablation study results showed that DFAM-DETR has adequate detection accuracy on small and medium

Table 5 The number of small, medium, and large slender objects within each category.

Categories	knife	spoon	book	fork	toothbrush
small	150	105	676	48	19
medium	123	99	366	87	26
large	53	49	119	80	12
Total	326	253	1161	215	57
Categories	snowboard	skis	surfboard	keyboard	baseball bat
small	23	103	69	10	32
medium	32	88	123	62	70
large	14	50	77	81	44
Total	69	241	269	153	146

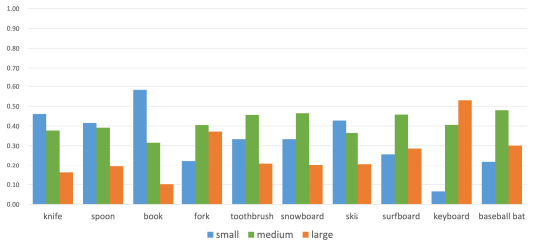


Fig. 10 The proportion of small, large, and large and medium slender objects within each category.

slender objects, except for large ones. A detailed analysis of the results is also conducted, see Table 5. Large slender objects, such as baseball bats and keyboards, have lower detection accuracy when using DFAM-DETR, see Fig. 10. The deformable convolution has the advantage of adapting its receptive field to the target object. The size of receptive field is sufficient to capture the majority of feature adaptively with deformable convolution for small slender objects. However, it cannot capture large slender objects in the last layer of the backbone network. The size of receptive field is insufficient and can only capture partial of the feature. As a result, the proposed DFAM-DETR is better at detecting small and medium slender objects than large ones.

5. Conclusion

This study proposes DFAM-DETR, a slender object detector based on Deformable DETR. Comparing to other popular detectors, it delivers greater detection accuracy for slender objects. With the proposed DFAM's deformable convolution and attention mechanism, it overcomes the limitation of convolution with fixed sampling points for slender object detection. DFAM-DETR detector improves the feature extraction ability of slender objects with greater detection accuracy and convergence speed. The research of DFAM-DETR will be expanded to include its detection capability and performance on various size of slender objects.

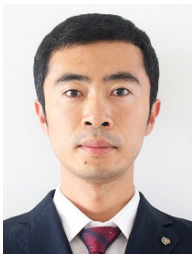
Acknowledgments

This work was supported by the 2020 Program for Liaoning Excellent Talents (LNET) in University, the 2021 Shenyang Ligong University Research Team Innovation Project, SYLUTD202105, and the Research Support Program for

Inviting High-Level Talents grant of Shenyang Ligong University (1010147000825).

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE conference on computer vision and pattern recognition, pp.770–778, 2016.
[2] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," IEEE access, vol.7, pp.128837–128868, 2019.
[3] Y. Zhao, Y. Rao, S. Dong, and J. Zhang, "Survey on deep learning object detection," Journal of Chinese imagegraphics, vol.25, no.4, p.26, 2020.
[4] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," Advances in Neural Information Processing Systems, vol.29, 2016.
[5] A. Bochkovskiy, C.Y. Wang, and H.Y.M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.
[6] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," Proc. IEEE conference on computer vision and pattern recognition, pp.7263–7271, 2017.
[7] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Proc. IEEE conference on computer vision and pattern recognition, pp.779–788, 2016.
[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg, "Ssd: Single shot multibox detector," European conference on computer vision, pp.21–37, Springer, 2016.
[10] C.Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A.C. Berg, "Dssd: Deconvolutional single shot detector," arXiv preprint arXiv:1701.06659, 2017.
[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," Proc. IEEE international conference on computer vision, pp.2980–2988, 2017.
[12] M. Tan, R. Pang, and Q.V. Le, "Efficientdet: Scalable and efficient object detection," Proc. IEEE/CVF conference on computer vision and pattern recognition, pp.10781–10790, 2020.
[13] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," Proc. IEEE/CVF international conference on computer vision, pp.9627–9636, 2019.
[14] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," Proc. European conference on computer vision (ECCV), pp.765–781, 2018.
[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proc. IEEE conference on computer vision and pattern recognition, pp.580–587, 2014.
[16] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," Proc. IEEE conference on computer vision and pattern recognition, pp.6154–6162, 2018.
[17] R. Girshick, "Fast r-cnn," Proc. IEEE international conference on computer vision, pp.1440–1448, 2015.
[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in Neural Information Processing Systems, vol.28, 2015.
[19] B. Zoph, E.D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q.V. Le, "Learning data augmentation strategies for object detection," European conference on computer vision, pp.566–583, Springer, 2020.
[20] L. Huang, Y. Yang, Y. Deng, and Y.D. Yu, "Unifying landmark localization with end to end object detection," arXiv preprint arXiv:1509.04874, 2015.
[21] P.R. Florence, L. Manuelli, and R. Tedrake, "Dense object nets: Learning dense visual object descriptors by and for robotic manipulation," arXiv preprint arXiv:1806.08756, 2018.
[22] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," Proc. IEEE/CVF International Conference on Computer Vision, pp.9657–9666, 2019.
[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft coco: Common objects in context," European conference on computer vision, pp.740–755, Springer, 2014.
[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," European conference on computer vision, pp.213–229, Springer, 2020.
[26] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," International Conference on Learning Representations, 2021.
[27] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," Proc. IEEE international conference on computer vision, pp.764–773, 2017.
[28] Z. Wan, Y. Chen, S. Deng, K. Chen, C. Yao, and J. Luo, "Slender object detection: Diagnoses and improvements," arXiv preprint arXiv:2011.08529, 2020.
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems, vol.30, 2017.
[30] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., "A survey on visual transformer," arXiv e-prints, pp.arXiv–2012, 2020.
[31] Y. Zhang, X. Shi, S. Mi, and X. Yang, "Image captioning with transformer and knowledge graph," Pattern Recognition Letters, vol.143, pp.43–49, 2021.
[32] C. Yang, Q. Wang, J. Du, J. Zhang, C. Wu, and J. Wang, "A transformer-based radical analysis network for chinese character recognition," 2020 25th International Conference on Pattern Recognition (ICPR), pp.3714–3719, IEEE, 2021.
[33] W. Liu, Y. Song, D. Chen, S. He, Y. Yu, T. Yan, G.P. Hancke, and R.W. Lau, "Deformable object tracking with gated fusion," IEEE Trans. Image Process., vol.28, no.8, pp.3766–3777, 2019.
[34] Z. Liu, B. Yang, G. Duan, and J. Tan, "Visual defect inspection of metal part surface via deformable convolution and concatenate feature pyramid neural networks," IEEE Trans. Instrum. Meas., vol.69, no.12, pp.9681–9694, 2020.
[35] J. Chen, Y. Chen, W. Li, G. Ning, M. Tong, and A. Hilton, "Channel and spatial attention based deep object co-segmentation," Knowledge-Based Systems, vol.211, p.106550, 2021.
[36] H.W. Kuhn, "The hungarian method for the assignment problem," Naval research logistics quarterly, vol.2, no.1-2, pp.83–97, 1955.
[37] K. Da, "A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[38] J. Deng, "A large-scale hierarchical image database," Proc. IEEE Computer Vision and Pattern Recognition, 2009.
[39] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, vol.60, no.6, pp.84–90, 2017.



Feng Wen received his B.S. degree from the Jilin University of Technology, China, in 1999, and his M.S. degree from Northeast University, China, in 2005. He received his Ph.D. degree from Waseda University, Japan, in 2010. He is currently a professor at the Graudate College, Shenyang Ligong University, China.

Mei Wang received her B.S. degree from the Qingdao University of Technology, China, in 2020, she is currently a postgraduate student at Shenyang Ligong University. Her research interests include artificial intelligence, evolutionary algorithms and applications.

Xiaojie Hu received his B.A. degree from the Portsmouth University, UK, and his M.A. degree from Western Michigan University, USA, in 2011. He received his Ph.D. degree from Western Michigan University, USA, in 2018. He is currently a lecturer at School of Information Science and Engineering, Shenyang Ligong University, China.