Design Studies and Intelligence Engineering
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA220704

Rotated Object Detector with Self-Attention and Improved IoU Loss

Qian Haiyang [a], Li Ning [a], Yuan Ning [a], Zhang Zhengran [a], and Zhou Huiyu [b]
[a] *Nanjing University of Aeronautics and Astronautics, Nanjing, P R China*
[b] *University of Leicester, Leicester, UK, LE1 7RH*

Abstract. Nowadays in UAV and remote sensing image processing area, rotated object detection has been attached more and more attentions. However, there are few studies of this based on Transformer, which is much stronger than CNN for feature extraction. Moreover, the well-developed IoU based loss function in horizontal object detection area is not well fits with rotated objects. In this paper, we argue that Transformer of pyramid structure called Swin-Transformer is an effective alternative of CNN. Specifically, the finetuned Swin-Transformer with Relatively Position Encoding (RPE) performed much better than other backbones generally used. Moreover, the new kind of IoU loss called Gaussian Estimate Loss (GE Loss) that use gaussian kernel to model object is applied in our model. It can increase the precious of the model. This loss is in contract to other Gaussian modeling loss function for adding direction vector that can solve the difficulty of testing objects close to square. Experiments on DOTA dataset achieved 84.99% map, which indicates that and our experiment shows that these improvements of our model are effective.

Keywords. Rotated Object Detection, Transformer, Position Encoding, IoU Loss

1. Introduction

As the importance of aerial photography and military increasing, remote sensing image and UAV image processing has been paid more and more attention. Large aspect ratio, densely arrangement and rotated bounding box are the main characteristics of these bird's eye view images. A horizontal bounding box would contain the rotated objects together, which would cause the difficulty in accurate detection and classification. An obvious solution is to add a parameter θ which represent the rotated angel for regression in traditional methods. However, general detectors is inefficient for the progress of gen- erating rotated bounding boxes and doing regression, let alone other problems with the angle that can dramatically decrease the precious. Specific detectors for rotated object should be designed.

Researchers have proposed a series of detector speccially designed for rotated ob- jects such as Rotated RetinaNet[6], R3Det[13], S2ANet[4], Rotated Reppoints[18], Ro- tated FasterRCNN[8], ROI Transformer[2] and Oriented RCNN[11]. However, prob- lems occured in these detectors as well. Firstly, their backbone for feature extraction is mainly based on CNN, whose ability of gaining features is not strong enough especially in rotated object detection tasks. For example, original R3Det requires some added feature refinement network to achieve higher precision. Secondly, the IoU loss function

usually used in horizontal object detection cannot be directly applied for that the area of the intersection of two rotated rectangles cannot be described by existing operators.
To solve the problems, Rotated-object Detector with Self-Attention (SARD) network is proposed. For feature extraction, self-attention is applied in the backbone to replace convolution. And RPE is used in self-attention to adapt rotated object better. For rotated bounding box IoU loss function, GE Loss based on kfIoU is imported to increase the accuracy of the area of intersecting regions in many cases.

2. Related Work

2.1 R3Det Rotated Objects Detectors

R3Det is a detector designed for rotated objects. The structure is shown in Figure 1. The Class & Box subnet is to do bounding box regression and classification. The main difference of the detector between other detectors is the feature refinement blocks. They can be multipled by many times and after many times of refinement, the Class & Box subnet followed can deal with more detailed features to improve the performance of the network.

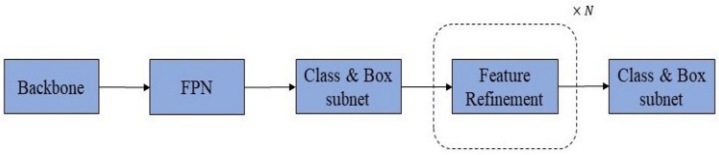
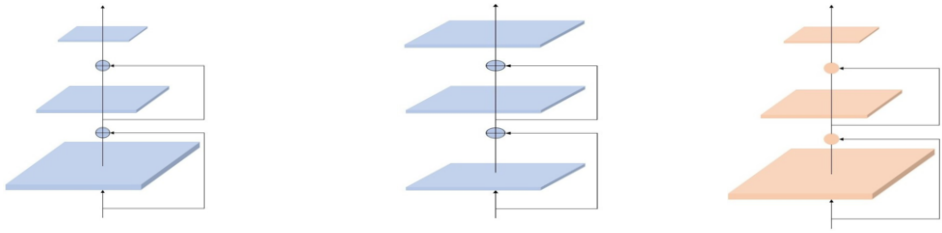


Figure 1. R3Det Detector

2.2 Self-Attention Mechanism and Swin-Transformer

Self-attention is firstly proposed in Transformer [10]. Instead of convolution, it brings an operator named "self-attention" which use three matrix to modeling an input sequence. This network can be much higher in precious than CNN in horizontal object detection[1, 19] and classification[9,3].
In the application of Transformer in subsequent tasks, a phenomenon is very com- mon that a specific Transformer is required for a specific task because the great calcula- tion and the difficulties in training. Nowadays, there is few works about specific designed Transformer for rotated object detection.
Swin-Transformer [7] is the short name of Shift-window Transformer. It is a general architecture proposed to solve two main problems of Transformer. One is poor network versatility and the other is difficulty in training. Swin draws on the mechanism of the convolution kernel of CNN, uses the sliding window self-attention method, performs self-attention calculation in a certain size window, and expands the receptive field through the movement of the window.
In contrast to other Transformer based detectors, Swin has a pyramid structure which gives it ability to extract muti-scale features, which is the same as CNN, shown in Figure 2. This character makes it possible for Swin to adapt different downstream tasks.



(a) Pyramid structure of CNN

(b) Columnar structure of Transformer

(c) Swin-Transformer has the same structure as CNN

Figure 2. Comparision of the structures

3. Proposed Method

The overview of our method SARD is shown in Figure 3. Finetuned Swin-Transformer is applied in the network as a backbone and RPE is used as well. A new loss function GE Loss is used in Class & Box subnet.

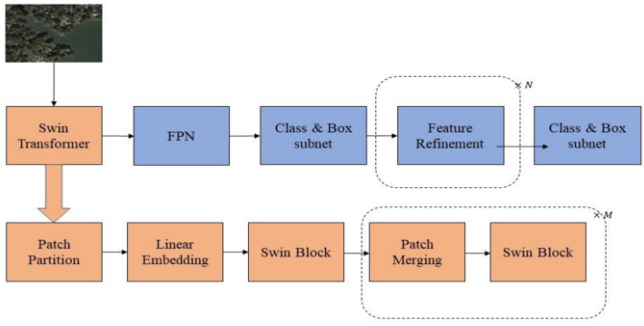


Figure 3. Structure of our detector SARD

Position encoding is required to express the order of the input vector for that self-attention operator does not include information of position. Transformer use Absolute Position Encoding (APE) to solve the problem. By defining a vector that present the position for every pixel and directly adding the vector to the input image, the self-attention operator can contain the position information. However, when it comes to rotate object detection, APE cannot present the relationship of each pixel.

3.1. Finetuned Swin-Transformer Backbone

The main part of the Swin Block is Window-Multihead-Self-Attention(W-MSA) modules and Shift-Window-Multihead-Self-Attention (SW-MSA) modules. W-MSA calculate the self-attention of each window and SW-MSA shift the window then calculate the self-attention. They are used alternately in a Swin Block. We found that changing all the W-MSA module to SW-MSA module, as shown in Figure 4, can give a faster convergence, shown as Table 1. The finetuned Swin-Transformer is applied to extract features for our detector.

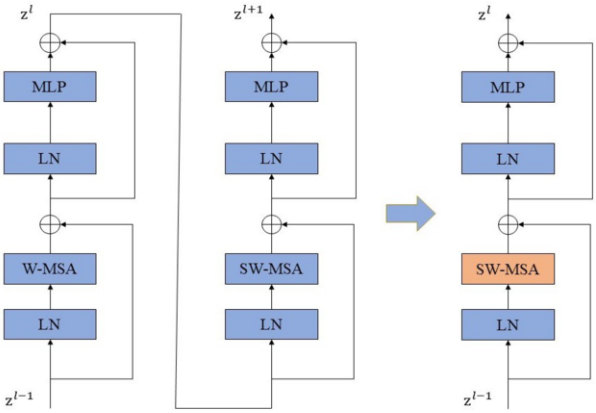


Figure 4. Change of Swin Block

Table 1. Change of the speed of convergence when substituting the module in Swin Block

Module in Swin Block	Training epoches	mAP
W-MSA and SW-MSA	12	77.89%
	24	80.03%(+2.14%)
SW-MSA	12	79.57%
	24	80.12%(+0.45%)

3.2. Relatively Position Encoding

RPE can represent the relationship of two pixels. In Figure 5. we take a 2 × 2 image as an example. Firstly, we mark a pixel as an anchor and get an encoding table. Then turn it into a row vector. Adding an offset to every dimension of the vector so that there are no minus number exist. Finally multiply 2 × *size* (*size of the window*) to the first dimension of each pixel and add the two dimensions together to get a number. By attaching the vector of every pixel together, the 4 × 4 tensor is the encoding table.

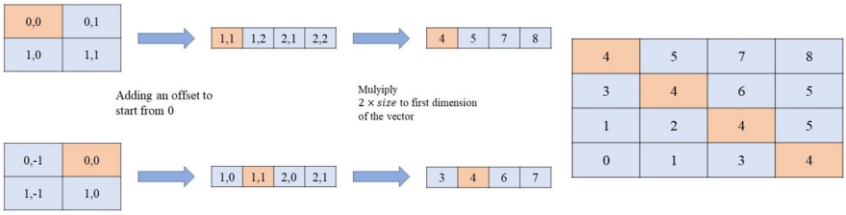


Figure 5. Process of RPE

3.3. Loss Function for Rotated Object

In object detection, IoU loss function is generally used to represent the difference of the predict bounding box and the ground truth. IoU loss can be present by Eq. (1)

$$L_{IoU}=1-\left|\frac{B_{pre}\cap B_{gt}}{B_{pre}\cup B_{gt}}\right| \tag{1}$$

In horizontal object detection, IoU loss is very practical. However, when it comes to rotating object detection, rectangle bounding box becomes very difficult for to calculate. And, the IoU loss of two rectangle bounding box, which is called skewIoU, is not derivable.

A solution to the problem is to turn the bounding box into gaussian kernel. By applying gaussian kernel, the characters of the bounding box can be presented by the parameters of the Gaussian distribution, which is easy to do further calculation with existing operators [17]. However, when the rectangle bounding box is close to a square, the gaussian kernel is close to a cycle instead of an ellipse. As the cycle is rotational symmetry, the rotated angle of the gaussian kernel is difficult to estimate. So, a direction vector is added to solve the problem. Figure 6 shows how to turn a rectangle bounding box to a gaussian kernel.

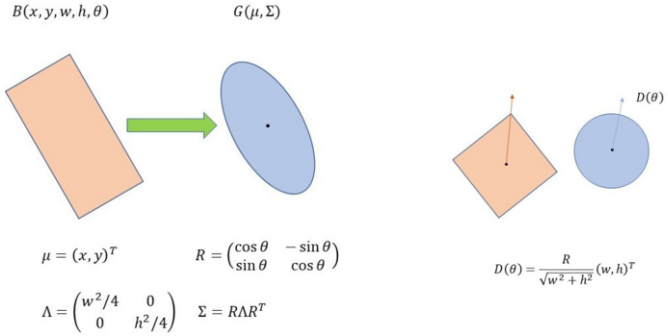


Figure 6. The process of Turing a bounding box to a gaussian kernel

KFIoU[17] is used to to indicates the area of the intersection of the Gaussian kernels. It can be expressed by Eq. (2)

$$\text{KFIoU}=\frac{V_{B3}(\Sigma)}{V_{B1}(\Sigma)+V_{B2}(\Sigma)-V_{B3}(\Sigma)} \tag{2}$$

For

$$V_B(\Sigma)=2^n\sqrt{\prod \text{eig}(\Sigma)} \tag{3}$$

$V_B(\Sigma)$ indicates the cumulative multiplication of the eigenvalues of covariances. Since the eigenvalues of covariance corresponds to the two sides of the bounding box, the cumulation of it is the area of the bounding box. Viewing the intersection of the boxes as a gaussian kernel the same as the bounding boxes, which we can easily get from the product of the gaussian kernels of the two bounding boxes. Then we can calculate the area of it by using the covariance.

However, if only multiplying the eigenvalues of covariance of the two gaussian kernel together, wherever the two gaussian kernels are, the result is the same. It is unreasonable because if the two bounding box is disjoint, the area of the intersection should be 0 because the covariance does not include the position relationship of the gaussian kernels. The mean of the gaussian kernel present the coordinate of the midpoint of the gaussian kernel. Thus, mean alignment is required before using covariance to calculate the area

of intersection. That is, adding the modulus of the mean difference to the loss. With the convergence of loss function, the modulus finally comes to 0 to make the gaussian kernel move together.

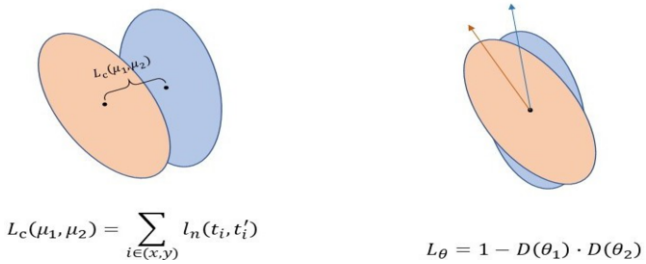


Figure 7. Mean alignment and direction correction

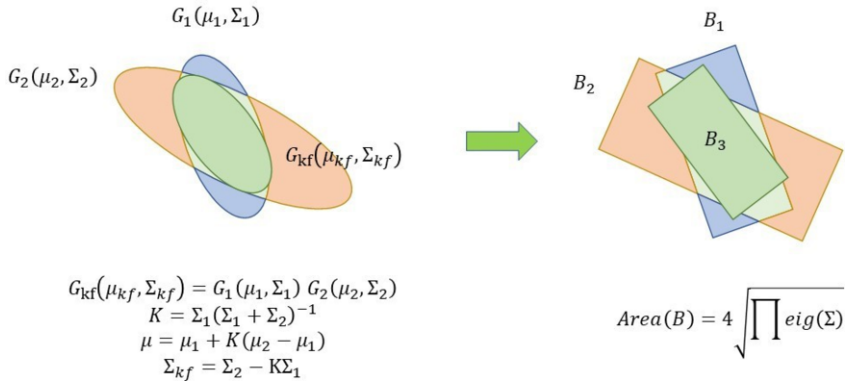


Figure 8. Turn gaussian kernel back to bounding box to measure the area of the intersection

The loss function, namely Gaussian kernel Estimation loss (GE-loss) can be presented as followed

$$L_{GE} = (1 - KFIoU) + L_c + L_\theta \tag{4}$$

The three terms of the equation represent the KFIoU, the distance of the center point and the difference in direction, or the difference in rotated angle.

4. Experiment

4.1. Details on Training

The experiment is on DOTA1.0 dataset. It is a public dataset which contains 15 classes of 2,806 large aerial images from different sensors and platforms. The dataset is divided into training, validation and testing sets. The model is trained on the training set with the help of some data augment methods such as random flip, random resize and color jitter. Random rotate are also used in the models. The platform contains 4 Titan v100 GPUs. For each model, the training is total 12 epochs, the original learning rate is 10^{-4} and after 8 epochs of training, it drops to 10^{-8}. The test is proceeding on the validation set.

4.2. Ablation Study

R3Det is the baseline of the experiment. It shows that the SARD network can greatly improve the performance. Changing the backbone to Swin with RPE can improve the mAP by 8.41% while the Swin with APE only improve the mAP by 3.48%. It shows that the self-attention operation can do better than traditional CNN in feature structure. The improvement of loss function is also efficient. The KF-IoU can improve the mAP by 4.9% alone and on the basic of it, the KF-IoU with direction vector can improve the mAP by 1.57%. The combination of Swin an KF-IoU shows a very high mAP over 80%. The number of features refinement blocks were adjusted in the experiment. The reduction of feature refinement blocks in original detectors caused the decrease of mAP and extra blocks made a little improvement. But at the same time, the adjustment of feature refinement blocks made little effect on the SARD model. The result is shown in the following Table 2.

4.3. Comparison with Swort-Of-The-Art

Comparison of our method with other method on DOTA1.0 dataset is shown in Table 3. The result shows that our method is better on mAP than other method include the two-stage method Oriented RCNN and ROI Transformer and Redet with equivariant convolution. SARD can achieve 84.99% mAP on DOTA1.0 dataset, which is very remarkable.

Table 2. Results of Ablation Study

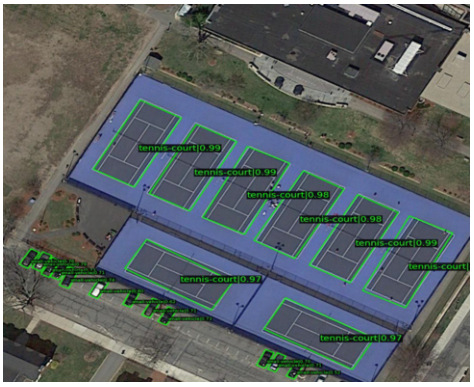
Method	Back bone		Loss Function		Feature refinement blocks	mAP
	Swin-APE	Swin-RPE	KF-IoU	GE Loss		
R3Det					2	71.16%
					4	71.49%
		✓			2	76.06%
			✓		2	76.64%
				✓	2	78.21%
			✓		4	77.03%
				✓	4	78.68%
SARD	✓				2	74.62%
	✓		✓		2	76.95%
	✓			✓	2	77.53%
	✓			✓	4	77.85%
		✓			2	79.57%
		✓	✓		2	82.01%
		✓		✓	2	84.99%
		✓		✓	4	84.96%
		✓		✓	6	84.84%

Table 3. Comparison with the Swort-Of-The-Art

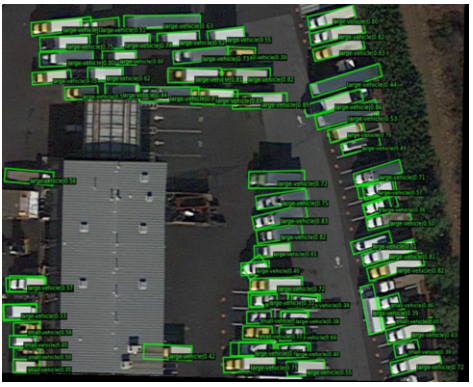
Method	mAP
RetinaNet-OBB[6]	**69.04%**
Rotated Reppoints[18]	**69.58%**
CSL[14]	**69.22%**
S2ANet[4]	**78.09%**
GWD[15]	**77.75%**
KLD[16]	**77.97%**
Gliding Vertex[12]	**77.37%**
Rotated Fast-RCNN[8]	**79.18%**
ROI-Transformer[2]	**82.51%**
Oriented RCNN[11]	**81.20%**
ReDet[5]	**79.94%**
SARD Net(Ours)	**84.96%**

4.4. Visualization Results

Here are some results tested by our model. In the picture, the large targets are easy to get a high probability of prediction. But the results of some very small targets are not so well. It might be improved by adding muti-scale training in our model. In (a), large objects such as tennis court are easy to be detected. And in (d), planes got a high degree of confidence though they are irregular in shape and near to square in bounding box. However, small objects in (b) and (c) show a degree of confidence not high enough, which needs further improvement. There is few conditions when the detector have miss detection or false alarm. Also, the densely arranged objects can be separated. In general, the performance is satisfactory.



(a)



(b)

(c) (d)

Figure 9. visualization results tested by the proposed model

5. Conclusion

In this paper, we discussed the weakness in existing rotated object detection methods, and put forward our solution, which is called SARD. Firstly, self-attention network Swin-Transformer is applied to take the place of CNN as the backbone. Relatively position encoding, which shows a better performance than former absolute position encoding is used in the backbone and all W-MSA model is changed by SW-MSA to simplify the model. Secondly, to describe the rotated rectangle bounding boxes better, gaussian kernel is used to modeling the bounding boxes. Based on KFIoU, GE Loss is proposed. Direction vector and mean alignment are added in our loss function. We do our experiment on DOTA dataset and it shows that our model can achieve near 85% mAP on our test.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *ArXiv*, abs/2005.12872, 2020.

[2] J. Ding, N. Xue, Y. Long, G. S. Xia, and Q. Lu. Learning roi transformer for oriented object detection in aerial images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

[4] J. Han, J. Ding, J. Li, and G. S. Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1–11, 2021.

[5] Jiaming Han, Jian Ding, Nan Xue, and Guisong Xia. Redet: A rotation-equivariant detector for aerial object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2785–2794, 2021.

[6] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PP(99):2999–3007, 2017.

[7] 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*, abs/2103.14030, 2021.

[8] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6):1137–1149, 2017.

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. In *ICML*.

[10] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,

Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.
[11] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3500–3509, 2021.
[12] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Guisong Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:1452–1459, 2021.
[13] Xue Yang, Qingqing Liu, Junchi Yan, and Ang Li. R3det: Refined single-stage detector with feature refinement for rotating object. In*AAAI*.
[14] Xue Yang, Junchi Yan, and Tao He. On the arbitrary-oriented object detection: Classification based approaches revisited. *Int. J. Comput. Vis.*, 130:1340–1365,2022.
[15] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. Rethinking rotated object detection with gaussian wasserstein distance loss. In *ICML*.
[16] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learn- ing high-precision bounding box for rotated object detection via kullback-leibler divergence. *ArXiv*, abs/2106.01883, 2021.
[17] Xue Yang, Yue Zhou, Gefan Zhang, Jitui Yang, Wentao Wang, Junchi Yan, Xiaopeng Zhang, and Qi Tian. The kfiou loss for rotated object detection. *ArXiv*, abs/2201.12558, 2022.
[18] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin. Reppoints: Point set representation for object detection. In 2019 *E/CVF International Conference on Computer Vision (ICCV)*.
[19] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: De- formable transformers for end-to-end object detection. *ArXiv*, abs/2010.04159, 2021.