# Time-rEversed diffusioN tEnsor Transformer: A new TENET of Few-Shot Object Detection (Supplementary Material)

Shan Zhang[⋆,†], Naila Murray[♣], Lei Wang[♦], and Piotr Koniusz[⋆,§,†]

[†]Australian National University [♣]Meta AI
[♦]University of Wollongong [§]Data61/CSIRO
[†]firstname.lastname@anu.edu.au, [♦]leiw@uow.edu.au, [♣]murrayn@fb.com

Below are additional derivations, evaluations and illustrations of our method.

## A Ablation Study on Encoding Network

Below we perform ablations of the backbone (Encoding Network, termed as EN in main paper). We use ConvNet (ResNet-50) and Transformer network [54] (Swin-B$^7$/ Swin-B$^{12}$ pre-trained on ImageNet-22K [5] with window size of 7/12), as shown in Table 5c. The comparisons are conducted by changing the backbone, whereas other settings remain unchanged. When ResNet-50 is replaced by Swin-B$^7$, we gain an improvement of 0.3% and 0.5% in the 5/10-shot setting (novel classes).

## B Details of Transformer Relation Head (TRH) with Z-shot and Spatial-HOP blocks.

As Z-shot T-RH is described in Eq. (15) of the main paper, below we focus on describing Spatial-HOP T-RH.

This head first forms a so-called self-attention on a set $\boldsymbol{\mathcal{Z}}$ of support regions and $\boldsymbol{\mathcal{B}}$ query RoIs, respectively. We formulate its operation for $B$ query RoIs (refer §4.2 of main paper for support regions). Spatial-HOP T-RH takes as input RoI features $\{\boldsymbol{\Phi}_b^* \in \mathbb{R}^{2d\times N}\}_{b\in\mathcal{I}_B}$ ($2d$ because layer 5 of ResNet-50 maps $d$-dimensional features to $2d$-dimensional features) and $\{\boldsymbol{\psi}_b^* \in \mathbb{R}^d\}_{b\in\mathcal{I}_B}$. We split $\boldsymbol{\Phi}_b^*$ along the channel mode of dimension $2d$ to create two new matrices $\boldsymbol{\Phi}_b^{*u} \in \mathbb{R}^{d\times N}$ and $\boldsymbol{\Phi}_b^{*l} \in \mathbb{R}^{d\times N}$ for $b \in \mathcal{I}_B$. We let $\boldsymbol{\Phi}_b^{*l} = [\boldsymbol{\phi}_{b,1}^{*l}, ..., \boldsymbol{\phi}_{b,N}^{*l}] \in \mathbb{R}^{d\times N}$. Self-attention is then performed over $\boldsymbol{\mathcal{T}}_b$ containing vectors, in parallel across $B$ RoIs, *ie.*, $\{\boldsymbol{\mathcal{T}}_b\}_{b\in\mathcal{I}_B}$:

$$\boldsymbol{\mathcal{T}}_b = [\boldsymbol{\phi}_{b,1}^{*l}, ..., \boldsymbol{\phi}_{b,N}^{*l}, \bar{\boldsymbol{\phi}}_b^{*u}, \boldsymbol{W}_g\boldsymbol{\psi}_b^*], \tag{17}$$

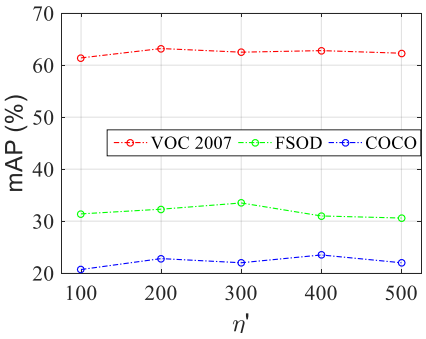
where $\bar{\boldsymbol{\phi}}$ denotes average-pooled features (FO) and $\boldsymbol{W}_g \in \mathbb{R}^{d\times d}$ denotes a linear projection (shared between query and support representations).

Based on these representations passed through the transformer head (variables indicated by widehat $\widehat{\cdot}$) between support regions and query RoIs, we then

Table 5: Experimental results of different variants of Transformer Relation Head (TRH), by varying Z-shot and Spatial-HOP blocks, are in Tab. 5a. Digits $1, ..., 4$ indicate different orders included or excluded from each experiment. "Spatial" is the size of spatial map (downsampled by the bilinear interpolation). Next, Tab. 5c is an ablation of different variants of Encoding Network (5/10-shot setting on VOC2007 testing set was used in Tab. 5a and 5c). Finally, Fig. 5b shows mAP w.r.t. $\eta'$ in SigmE (10-shot protocol on VOC2007 and COCO testing dataset, 5-shot setting on FSOD testing dataset).

| $Z-$shot (1,2,3,4) | Spatial | 1 | 2,3,4 | 5-shot | 10-shot |
|---|---|---|---|---|---|
| | 7×7 | ✓ | | 57.9 | 64.2 |
| | | | ✓ | 61.3 | 65.8 |
| | | ✓ | ✓ | **62.3** | **66.9** |
| | 5×5 | ✓ | | 58.7 | 63.7 |
| ✓ | | | ✓ | 60.3 | 64.3 |
| | | ✓ | ✓ | 61.1 | 65.2 |
| | 3×3 | ✓ | | 54.8 | 57.9 |
| | | | ✓ | 56.0 | 59.2 |
| | | ✓ | ✓ | 56.6 | 60.1 |
| | 1×1 | ✓ | | 45.1 | 49.3 |
| | | | ✓ | 46.8 | 51.9 |
| | | ✓ | ✓ | 47.4 | 52.4 |
| | 7×7 | ✓ | | 55.2 | 60.7 |
| | | | ✓ | 59.4 | 64.6 |
| | | ✓ | ✓ | 61.0 | 65.8 |
| | 5×5 | ✓ | | 57.6 | 61.5 |
| | | | ✓ | 58.6 | 63.0 |
| | | ✓ | ✓ | 60.3 | 63.4 |
| | 3×3 | ✓ | | 52.4 | 54.8 |
| | | | ✓ | 54.1 | 57.8 |
| | | ✓ | ✓ | 54.5 | 58.3 |
| | 1×1 | ✓ | | 44.0 | 47.1 |
| | | | ✓ | 45.2 | 48.4 |
| | | ✓ | ✓ | 46.3 | 50.1 |

(a)




(b)

| EN | 5-shot | 10-shot |
|---|---|---|
| ResNet-50 | 62.3 | 66.9 |
| Swin-B$^{7}$ | **62.6** | **67.4** |
| Swin-B$^{12}$ | 62.0 | 66.7 |

(c)

compute relations as follows:

$$\mathcal{R}^b_{\text{Spatial}} = \left[\widehat{\boldsymbol{\Phi}}^{\dagger l} - \widehat{\boldsymbol{\Phi}}^{* l}_b\right] \in \mathbb{R}^{d\times N}, \quad b \in \mathcal{I}_B, \tag{18}$$

$$\mathcal{R}^b_{\text{FO+HO}} = \begin{bmatrix} \widehat{\boldsymbol{\phi}}^{\dagger u} \cdot \widehat{\boldsymbol{\phi}}^{* u}_b \\ \widehat{\boldsymbol{\psi}}^{\dagger} \cdot \widehat{\boldsymbol{\psi}}^{*}_b \end{bmatrix} \in \mathbb{R}^{2d}, \tag{19}$$

$$\mathcal{R}^b = \begin{bmatrix} \text{Repeat}(\mathcal{R}^b_{\text{Spatial}}; N) \\ \boldsymbol{W}^u \mathcal{R}^b_{\text{FO+HO}} \end{bmatrix} \in \mathbb{R}^{2d\times B}, \tag{20}$$

where the learnable weight $\boldsymbol{W}^{(u)} \in \mathbb{R}^{d\times 2d}$ projects the channel-wise concatenated matrix to $d$ dimensions, letters $l$ and $u$ indicate first and second half of channel coefficients, respectively, operator $\cdot$ indicates element-wise multiplication, and $\text{Repeat}(\cdot; N)$ replicates spatial mode $N$ times. The above process is shown in Fig. 4.3.

## C Ablation Study on Transformer Relation Head (TRH) with Z-shot and Spatial-HOP blocks.

As the supplementary setting for the top panel of Tab. 3b (in the main paper), we utilize $r=1$ in RPN and $r=2,3,4$ in TRH, achieving 2.7%/2.4% improvement on novel/base classes, 5-shot protocol, over the variant applied $r=1$ in both RPN and TRH.

We then conduct more ablation studies on Spatial-HOP transformer head to analyze the impact brought by each component (5/10-shot setting on novel classes, VOC 2007). The results are shown on Table 5a. Specifically, we mainly ablate three variants: spatial maps of assorted size (as in the table) with either orderless HOP representation of order $r=1$ or $r=2,3,4$, or both $r=1,2,3,4$.

Furthermore, to investigate the impact of spatial attention, we use bilinearly subsampled maps, ranging from $1\times 1$ to $7\times 7$ in spatial size. Not surprisingly, the Spatial-HOP head performs best when utilizing larger spatial maps, together with the orderless high-order and first-order tensor descriptors.

## D Visualization of Attention Maps of the Spatial-HOP block.

To explain why our model benefits from the combination of spatial attention, and orderless first-order and high-order representations, we provide qualitative results based on displaying attention maps.

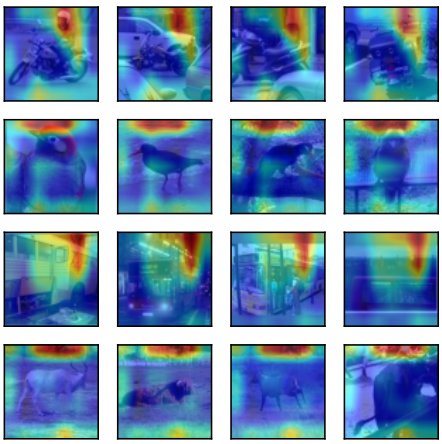
Firstly, we performed training where Spatial-HOP T-RH used only spatial and first-order information (FO) during training. To obtain the picture, we picked $\bar{\boldsymbol{\phi}}^{\dagger u}$ from Eq. (16) and we looked how it correlates with the $N$ spatial representations $\boldsymbol{\phi}_1^{\dagger l},...,\boldsymbol{\phi}_N^{\dagger l}$. To that end, we passed these "spatial fibers" and FO representation via the RBF kernel of Eq. (8), and we then reshaped $N$ into the spatial map ($7\times 7$ size).

Figure 3 (top left) shows how the first-order representation (FO) correlates with each spatial fiber in the attention of transformer. As Spatial-HOP T-RH block uses information averaged over $K$ images of the same class in an episode ($K$-way images), each column shows one of these support images. Each row shows a different class image from $Z$-shot support images in the episode.

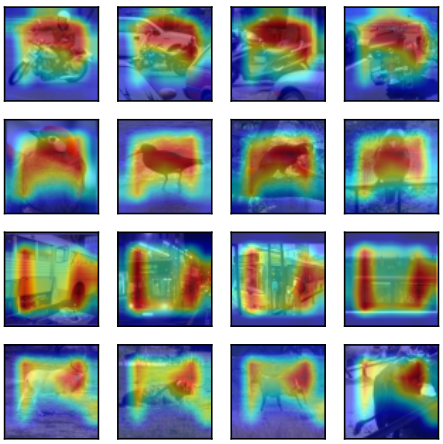
Subsequently, we performed training where Spatial-HOP T-RH used only spatial and high-order information (HO) during training. Thus, we picked the high-order representation $\boldsymbol{W}^{(g)}\boldsymbol{\psi}^{\dagger}$ from Eq. (16) and we looked how it correlates with the $N$ spatial representations $\boldsymbol{\phi}_1^{\dagger l},...,\boldsymbol{\phi}_N^{\dagger l}$. To that end, we passed these "spatial fibers" and HO representation via the RBF kernel of Eq. (8), and we then reshaped $N$ into the spatial map ($7\times 7$ size).

Figure 3 (top right) shows how the high-order representation (HO) correlates with each spatial fiber in the attention of transformer. As before, we visualise $K\times Z$ images from an episode given the $K$-way $Z$-shot problem.

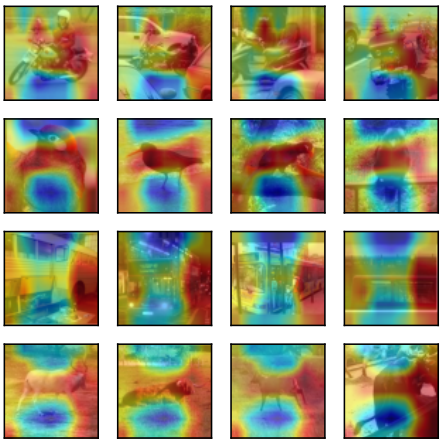
Comparing FO an HO representations, HO is by far more focused on the foreground objects that correlate in the semantic sense with the object class. This



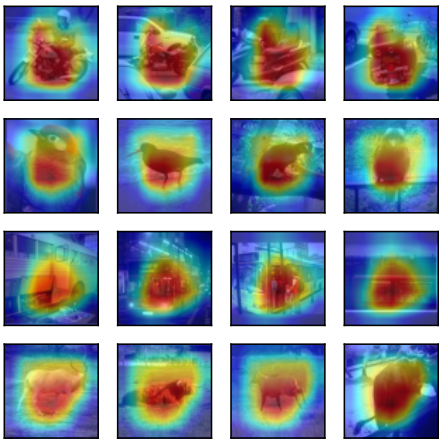
First-order fiber (FO) is visualised (Spatial-HOP T-RH used only spatial and FO ($r=1$) information during training)



High-order fiber (HO) is visualised (Spatial-HOP T-RH used only spatial and HOP ($r=2,3,4$) information during training)



Spatial fibers are max-pooled and then visualised (Spatial-HOP T-RH used spatial, FO and HOP information ($r=1,2,3,4$) during training)



First-order fiber (FO) and High-order fiber (HO) are averaged and then visualised (Spatial-HOP T-RH used spatial, FO and HOP information ($r=1,2,3,4$) during training)

Fig. 3: Visualization of attention maps of self-attention w.r.t. support regions. The results are produced on VOC2007 test set, novel classes (motorbike, bird, bus and cow). See text for detailed descriptions.

explains why HO representations help our model obtain better results compared to traditional attention mechanisms that focus only on capturing spatial correlations of a region.

Figure 3 (bottom left) shows how the spatial fibers from the attention matrix that is max-pooled along columns (we of course removed FO and HO before pooling along columns). We follow the same procedure as above, however, this time the Spatial-HOP T-RH block was utilizing the spatial, FO and HO information during training. Clearly, spatial attention can focus on complex spatial patterns in contrast to the focus of FO and HO.

Figure 3 (bottom right) shows how the first-order representation (FO), averaged with the high-order representation (HO), correlate with each spatial fiber in the attention of transformer. We follow the same procedure as above, and still use the spatial, FO and HO information in the Spatial-HOP T-RH block during training. Clearly, utilizing $r=1,2,3,4$ compares favourably with utilizing either $r=1$ or $r=2,3,4$ during training.

## E Impact of $\eta'$ of SigmE.

According to Section 4, TSO benefits from element-wise PN, realized by the SigmE operator in Eq. (5), which depends on the parameter $\eta'$. Figure 5b shows that $\eta'=200$ is a good choice on VOC dataset but $\eta'=300/400$ helps obtain the best results on FSOD/COCO dataset. Overall, our approach is not overly sensitive to this parameter, and setting $\eta'=200$ on all datasets if a good choice.

## F Hyperparameters on the FSOD and COCO datasets.

Table 6: Ablation studies on the FSOD and COCO datasets (5/10-shot, novel classes), w.r.t. the effect of varying (a) the number of heads used in T-Heads Attention, as shown in Tab. 6a, and (b) the number of TENET blocks as shown in Tab. 6b. mAP of variants of High-order Tensor Descriptors (HoTD) with TSO ($\eta_r>1$) and without TSO ($\eta_r=1$) is in Tab. 6c.

| $TA$ | FSOD | COCO |
|---|---|---|
| | 5-shot | 10-shot |
| 1 | 30.5 | 20.1 |
| 2 | **31.7** | 22.3 |
| 4 | 31.2 | 22.6 |
| 8 | 30.8 | **23.5** |
| 16 | 30.0 | 23.0 |
| 32 | 29.4 | 21.8 |
| 64 | 29.5 | 21.5 |

(a)

| $TB$ | FSOD | COCO |
|---|---|---|
| | 5-shot | 10-shot |
| 1 | 31.7 | 23.5 |
| 2 | **33.5** | 24.2 |
| 3 | 32.6 | **25.1** |
| 4 | 31.0 | 24.8 |
| 5 | 31.2 | 23.1 |

(b)

| $r$ | | | dim. split | $\eta_r$ (FSOD) | 5-shot | | $\eta_r$ (COCO) | 10-shot | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 3 | 4 | | | $AP_{50}$ | $AP_{75}$ | | $AP_{50}$ | $AP_{75}$ |
| ✓ | | | | 7 | 33.1 | 29.6 | 10 | 25.7 | 17.5 |
| ✓ | ✓ | | 3:1 | 7,7 | 33.7 | 30.4 | 10,10 | 26.0 | 18.2 |
| ✓ | ✓ | ✓ | 5:2:1 | 7,7,7 | **35.4** | **31.6** | 10,10,10 | **27.4** | **19.6** |
| ✓ | ✓ | ✓ | 5:2:1 | 1,1,1 | 30.8 | 28.4 | 1,1,1 | 22.1 | 14.3 |

(c)

Tables 6a and 6b present the impact of the number of head used in T-Heads Attention ($TA$) and TENET block ($TB$) on results. We fix the $\sigma = 0.5$ (the best value of standard deviation of the RBF kernel of transformers, selected by cross-validation on FSOD and COCO dataset) and then we investigate $TA$ and $TB$ (the number of attention units per block, and the number of blocks, respectively). Two heads together with two blocks are the best on the FSOD dataset, while eight heads aligned with three blocks yield the best results on the COCO dataset. Table 6c shows results on FSOD and COCO w.r.t. the dimension split along the feature channel (*e.g.*, if $r = 2, 3$, ratio 3:1 means that three parts of channel dimension are taken to form the second-order representation, and one part of channel dimension is taken to form the third-order representation). The table also shows the impact of $\eta_r$ of TSO on results, where $\eta_r$ are individual parameters for each order $r$. Overall, using all three orders, as denoted by $r = 2, 3, 4$, outperforms a second-order representation, indicated by $r = 2$. Importantly, TSO is used when $\eta_r > 1$. Without TSO ($\eta_r = 1$), results drop by a large margin, which highlights the practical importance of TSO on results.

## G Comparison with QA-FewDet/DeFRCN fine-tuning/meta-testing setting (Table 7).

Below we compare our method with with QA-FewDet [53]/DeFRCN [55].

Table 7: Comparison with QA-FewDet/DeFRCN (mAP%).

| Method | Encoding Network | Novel Set 1 | | | | | Novel Set 2 | | | | | Novel Set 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 | 1 | 2 | 3 | 5 | 10 |
| **Meta-training the model on base classes, and meta-testing on novel classes** | | | | | | | | | | | | | | | | |
| QA-FewDet | ResNet-101 | 41.0 | 33.2 | 35.3 | 47.5 | 52.0 | 23.5 | 29.4 | 37.9 | 35.9 | 37.1 | 33.2 | 29.4 | 37.6 | 39.8 | 41.5 |
| TENET (Ours) | ResNet-50 | **43.7** | **42.1** | **43.9** | **48.2** | **54.5** | **32.5** | **35.2** | **39.5** | **37.8** | **38.7** | **34.1** | **37.0** | **38.9** | **42.0** | **45.1** |
| **Fine-tuning the model on novel classes, and testing on novel classes** | | | | | | | | | | | | | | | | |
| QA-FewDet | ResNet-101 | 42.4 | 51.9 | 55.7 | 62.6 | 63.4 | 25.9 | 37.8 | 46.6 | 48.9 | 51.1 | 35.2 | 42.9 | 47.8 | 54.8 | 53.5 |
| DeFRCN | ResNet-101 | **53.6** | **57.5** | **61.5** | 64.1 | 60.8 | 30.1 | 38.1 | 47.0 | **53.3** | 47.9 | **48.4** | **50.9** | **52.3** | 54.9 | 57.4 |
| TENET (Ours) | ResNet-50 | 46.7 | 52.3 | 55.4 | 62.3 | 66.9 | 40.3 | 41.2 | 44.7 | 49.3 | 52.1 | 35.5 | 41.5 | 46.0 | 54.4 | 54.6 |
| TENET(Ours) | ResNet-101 | 48.5 | 55.2 | 58.7 | **65.8** | **69.0** | **42.6** | **43.4** | **47.9** | 52.0 | **54.2** | 37.9 | 43.6 | 48.8 | **56.9** | **57.6** |

## H Comparison with SOTA on MS COCO minival set (10/30 -shot) as shown in Table 8.

Table 8: Evaluations on the MS COCO minival set (10/30- shot). Methods that do not disclose all shot results are ignored and are replaced with '–'.

| Method | Venue | $AP$ | | $AP_{50}$ | | $AP_{75}$ | |
|---|---|---|---|---|---|---|---|
| | | 10 | 30 | 10 | 30 | 10 | 30 |
| FSCE+SVD | NeurIPS 2021 | 12.0 | 16.0 | – | – | 10.4 | 15.3 |
| FADI | NeurIPS 2021 | 12.2 | 16.1 | – | – | 11.9 | 15.8 |
| SRR-FSD | CVPR 2021 | 11.3 | 14.7 | 23.0 | 29.2 | 9.8 | 13.5 |
| Zhang *et al.* | CVPR 2021 | 12.6 | – | 27.0 | – | 10.9 | – |
| QA-FewDet | ICCV 2021 | 11.6 | 16.5 | 23.9 | 31.9 | 9.8 | 15.5 |
| DeFRCN | ICCV 2021 | 18.5 | 22.6 | – | – | – | – |
| QSAM | WACV 2022 | 13.0 | 15.3 | 24.7 | 29.3 | 12.1 | 14.5 |
| FCT | CVPR 2022 | 15.3 | 20.2 | – | – | – | – |
| TENET | Ours | **19.1** | **23.7** | **27.4** | **32.2** | **19.6** | **23.1** |

## I Mean ± std of mAP on PASCAL VOC 2007 (Table 9).

Table 9: Evaluations on three test splits of VOC 2007 (mean mAP ± std).

| Method/Shot | | Mean±std | | | |
|---|---|---|---|---|---|
| | | 1 | 3 | 5 | 10 |
| FRCN | ICCV12 | 7.6±3.1 | 23.5±4.5 | 32.3±3.3 | 36.4±6.0 |
| FR | ICCV19 | 16.6±1.9 | 25.0±1.7 | 34.9±4.3 | 42.6±3.4 |
| Meta | ICCV19 | 14.9±3.9 | 30.7±3.2 | 40.6±4.5 | 48.3±2.5 |
| FSOD | CVPR20 | 25.4±3.2 | 32.0±4.8 | 42.2±4.2 | 47.9±3.9 |
| NP-RepMet | NeurIPS20 | 37.6±3.4 | 41.6±1.5 | 45.4±2.8 | 47.8±2.1 |
| PNSD | ACCV20 | 31.3±4.4 | 36.2±4.2 | 44.5±3.8 | 49.9±5.4 |
| MPSR | ECCV20 | 33.9±7.2 | 44.3±5.2 | 47.7±6.2 | 53.1±6.2 |
| TFA | ICML20 | 31.4±6.7 | 40.5±4.6 | 46.8±8.6 | 48.3±7.0 |
| FSCE | CVPR21 | 31.4±9.3 | 44.8±4.9 | 51.1±7.7 | 55.9±5.6 |
| CGDP+FRCN | CVPR21 | 33.1±5.6 | 43.7±2.3 | 50.0±6.0 | 54.8±6.6 |
| TIP | CVPR21 | 24.0±2.6 | 38.4±4.0 | 45.2±4.3 | 52.5±5.3 |
| FSOD$^{up}$ | ICCV21 | 36.8±5.2 | 45.1±3.8 | 50.1±4.6 | 54.5±5.5 |
| QSAM | WACV22 | 26.1±3.5 | 35.4±2.9 | 45.4±3.6 | 51.9±3.8 |
| TENET | (Ours) | **40.8±3.6** | **48.7±4.7** | **55.3±3.1** | **57.9±5.8** |

## References

53. Han, G., He, Y., Huang, S., Ma, J., Chang, S.: Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. pp. 3243–3252. IEEE (2021)

54. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. CoRR **abs/2103.14030** (2021)
55. Qiao, L., Zhao, Y., Li, Z., Qiu, X., Wu, J., Zhang, C.: Defrcn: Decoupled faster R-CNN for few-shot object detection. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 8661–8670. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00856