Supplementary material for FS-DETR: Few-Shot DEtection TRansformer with prompting and without re-training

Adrian Bulat[1,2], Ricardo Guerrero[1], Brais Martinez[1], Georgios Tzimiropoulos[1,3]
[1]Samsung AI Cambridge [2]Technical University of Iasi [3]Queen Mary University of London

A. Implementations details

FS-DETR extends Conditional DETR [11] (see Section 3), and was pre-trained and trained on a single node with 8 P40 GPUs. Following [4], the ResNet50 [9] backbone is initialized from SwAV [3] and kept frozen. Pre-training makes use of ImageNet-100 without labels, with object proposals detection as a pretext task.

Pre-training hyper-parameters were set to: Batch size of 32 per GPU, AdamW optimizer [10] with a learning rate of 10^{-4}, frozen backbone CNN, path dropout of 0.1, training for 60 epochs with the learning rate decreased by factor of 10 after 40 epochs. When using larger images for pre-training (*i.e.* containing complex scenes) the batch size is decreased to 2.

Training hyper-parameters were set to: Batch size of 2 per GPU, SGD with momentum (0.9) [12] with the learning rate initially set to $5e^{-1}$, path dropout of 0.1, training for 30 epochs with the learning rate decreased by a factor of 10 after 20 epochs (and respectively 15 for COCO). Augmentation followed DETR: input images were resized such that the short axis is 480 at least and 800 pixels at most, and the long side is, at most, 1333 pixels, and randomly cropped with 0.5 probability.

Patch augmentation hyper-parameters. The templates are cropped tightly based on the bounding box and then rescaled to a 128×128px image. During training we apply the following augmentations: color jittering, with 0.8 probability and 0.4 intensity, random gray scale (0.2 probability) and Gaussian blur with a probability of 0.5.

B. Pre-training process

Transformer based architectures are known to generally be more data-hungry than their homologous CNNs [5, 2]. To alleviate this, we introduce a label-free pre-training step that closely mimics the training stage.

More specifically, at train time, for any given input image, we crop a set of patches according to the object proposals produced by Selective Search [17] [1]. Each of these patches represents an object (belonging to some class) and can be mapped to a pseudo-class, by associating it to a different pseudo-class embedding. Note, that random patches can be used too, but the former leads to faster convergence. The goal of the network is to predict the location of these patches (*i.e.* object templates). To make the task harder, the patches (templates) are augmented using a set of random transformations before being passed to the backbone. Finally, the network is trained using a regression (for the bounding boxes) and a classification loss. As opposed to the supervised training stage, the classification loss is reduced to a binary classification problem initially: object/no object and then to the proposed loss, after this warm-up. The model is then trained using the hyper-parameters described in Section A while the ResNet based backone is initialised from a model pre-trained on Imagenet without supervision (SwAV [3]). Note that unlike [4, 1] that also make use of unsupervised detection-centric training, our work concatentes a set of templates as prompts, instead of grouped-based summation, uses a differnt trainign objective and makes use of negative templates. The process is illustrated in Fig. 1.

Pre-training dataset For our DETR pre-training, we used the images belonging to the base classes from COCO (60 classes in total) and ImageNet-100 (a subset of ImageNet introduced in [16]). We note the following: firstly, there is no overlap between COCO base classes and VOC and COCO novel classes. Secondly, ImageNet-100 contains classes that can be matched to 7 out of 20 VOC classes (bird, cat, dog, boat, car, motorcycle and chair). Specifically, split-1 of VOC novel classes contains 2/5 classes (bird and motorbike) that overlap with ImageNet-100, split-2 0/5 and split-3 3/5 (boat, cat and motorbike). Please note that NONE of the labels in ImageNet-100 (or COCO) are used at any stage of the pre-training. While we agree that the underlying data distribution, even for unsupervised learning is important, judging from the results from Tables 1 and 2 the gains in absolute terms offered by our approach are consistent across all 3 sets (note that split-2 has no overlap at all).

We note that, recent state-of-the-art methods (*e.g.* Fan et al [6], QA-FewShot [7], DeFRCN [13]) make use of a

[1]Selective Search is a **training-free** generic region proposal algorithm that computes a hierarchical grouping of image regions based on color, texture, size and shape, and hence, has no notion of object classes.

backbone pre-trained with full supervision on the entire ImageNet, same which includes all VOC/COCO novel classes. In this regard, we trained FS-DETR initialized from a backbone pre-trained on the entirety of Imagenet for classification using full supervision (*e.g.* same as [6, 7, 13]). Preliminary results shown in Tab. 1 (which could likely be improved from hyper-parameter optimization) indicate an overall improvement of approx. 1.5%. This highlights that the pre-training data used in the proposed work doesn't offer any advantage over prior art approaches that use fully supervised pre-trained backbones. Further to this, DeFRCN [13] experimented with using a backbone pre-trained on ImageNet without labels (SwAV weights - same as ours) which resulted in substantially degraded performance of approx. 5.0%.

Table 1. Impact of different initialisation of backbone on the PASCAL VOC dataset (Novel Set 1).

Approach	Novel Set 1				
	1	2	3	5	10
FS-DETR (Swav)	45.0	48.5	51.5	52.7	56.1
FS-DETR (ImageNet)	47.1	49.9	52.5	53.8	57.0

C. Qualitative evaluation

Fig. 2 shows 1-shot detection examples of FS-DETR , with success cases shown on the first three columns, and fail cases on the right-most column. The image on top-left of the figure, illustrates an important and unique property of FS-DETR : Two novel classes coexist in a single image, and FS-DETR is able to successfully detect both of them at the same time.

Fig. 3 shows the effect of varying the 1-shot template used during novel class detection. There, smaller images refer to the templates used for 1-shot detection on the paired larger image. From the left-most two pairs of columns, it can be appreciated that even under large template visual variability, FS-DETR proves to be extremely robust, with detections hardly affected by the template change. The right-most illustrates a failure case, where the sofa fails to be detected.

Additionally, in Fig. 4 we visualise the attention weights between the visual prompts and the encoded image features. Notice that our network learns to attend to parts of the target image that are semantically similar to the provided templates that are present in the target image.

D. Discussion, challenges and limitations

Herein we offer a pertinent discussion on some things we tried but didn't work, defining some of the limitations and challenges that arise within the proposed framework and more so in general for FSOD using images within DETR framework.

D.1. Few-shot object detection objective ambiguity

A general limitation of few shot object recognition systems, trained and/or tested using one or more visual examples is the ill-definess of what represents a class. For example, presenting a template depicting a dog could require identifying the class "dog", "bulldog"(*i.e.* find dogs of a given bred), "a white dog" etc. While as the number of examples increases the ambiguity decreases, the problem is not fully solvable within the visual domain. A natural solution to this problem could be provided by constructing the templates using natural language. While an interesting solution, this goes beyond the scope of this work.

That being said, to some extent, our approach alleviates parts of this problem: As our model has to distinguish locally within the set of provided positive (present in the image) and negative (not present) templates, it can use them to semantically ground the notion of a class, effectively defining the semantic hierarchy. For example, if all templates are representing different apple varieties, the model is expected to differentiate between these varieties instead of detecting *any* apple.

D.2. Challenges within the DETR framework

Despite its remarkable success and appealing formulation that removes the need of an explicit object proposal component or post-processing step (*i.e.* NMS), in the context of few-shot detection some of this advantages pose additional challenges, some of which we detail bellow. We believe this aspects could represent potentially interesting future exploration directions.

Semantic misalignment Traditional object detection systems, such as [15, 14, 8] preserve an exact feature alignment between the regressed bounding box and the semantic information (*i.e.* the ROI pooling extracts features at the location given by the proposal). DETR derived approaches however construct their representation gradually by adapting a set of object queries via self-attention and cross-attention with the encoded features. As each object query operates (attends) to the entire image, as opposed to the local ROI, the query can encode information outside of the predicted bounding box. Thus, we can get to cases where the class may be correct although the bounding box contains mostly objects of an incorrect category.

Therefore, when we tried to use an external supervised classifier, applied to the image region cropped based on the predicted bounding box, surprisingly we noticed a deterioration of the performance. Upon visual inspection we observed a manifestation of the above mentioned phenomena, where the model was able to predict the correct class despite the fact that the predicted bounding box contained predominantly content of a different class, while the external supervised classifier was unable to.

Reduced proposal diversity A key characteristic of DETR

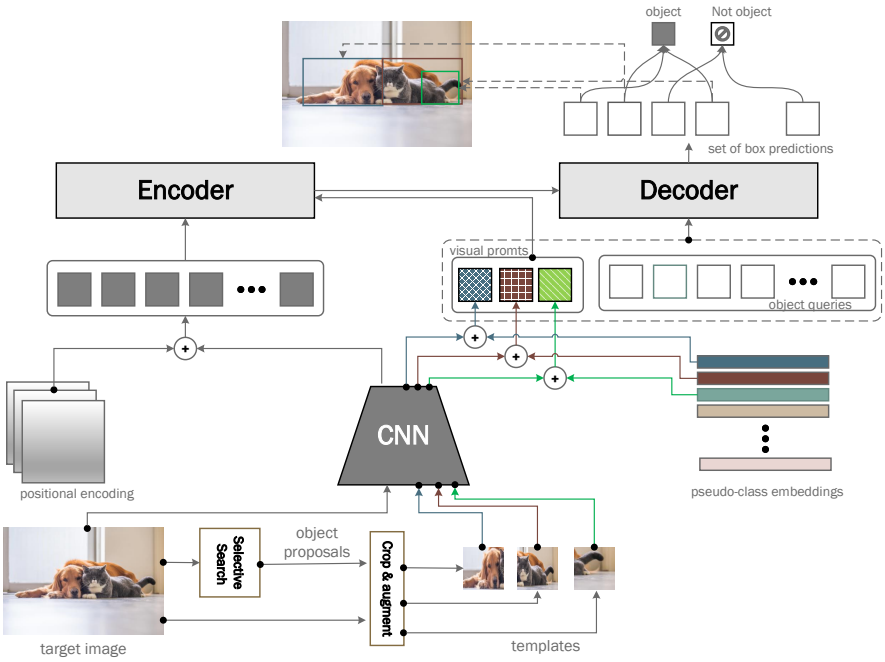

Figure 1. FS-DETR pre-training stage. The pre-training process largely mimics the training stage, with a few notable differences: (1) no annotations are used, (2) the target bounding boxes are proposed by selective search or sampled randomly, (3) the templates are sampled from the target image itself and (4) only two classes are defined - object and no object.

Figure 2. Novel class 1-shot detection examples with FS-DETR. First three columns depict success cases, while the right-most column failures. Green and red boxes indicate novel and base classes, respectively. Note that in the top-left image two novel classes are detected simultaneously.

systems is the removal of an a) external object proposal generator and b) implicit Non Maximum Suppression (NMS). Upon close inspection of our system we noticed that as we advance within the transformer based decoder, the bounding boxes are pruned via self-attention. By the end, despite having 100-300 object queries, most will point to a very small set of distinct regions of the image, lacking the diversity present in more traditional systems, such as in Fast RCNN. The consequence of this is a higher likelihood of missing unseen classes in limited data scenarios, making the pre-training even more so important to train the built-in object proposals system.

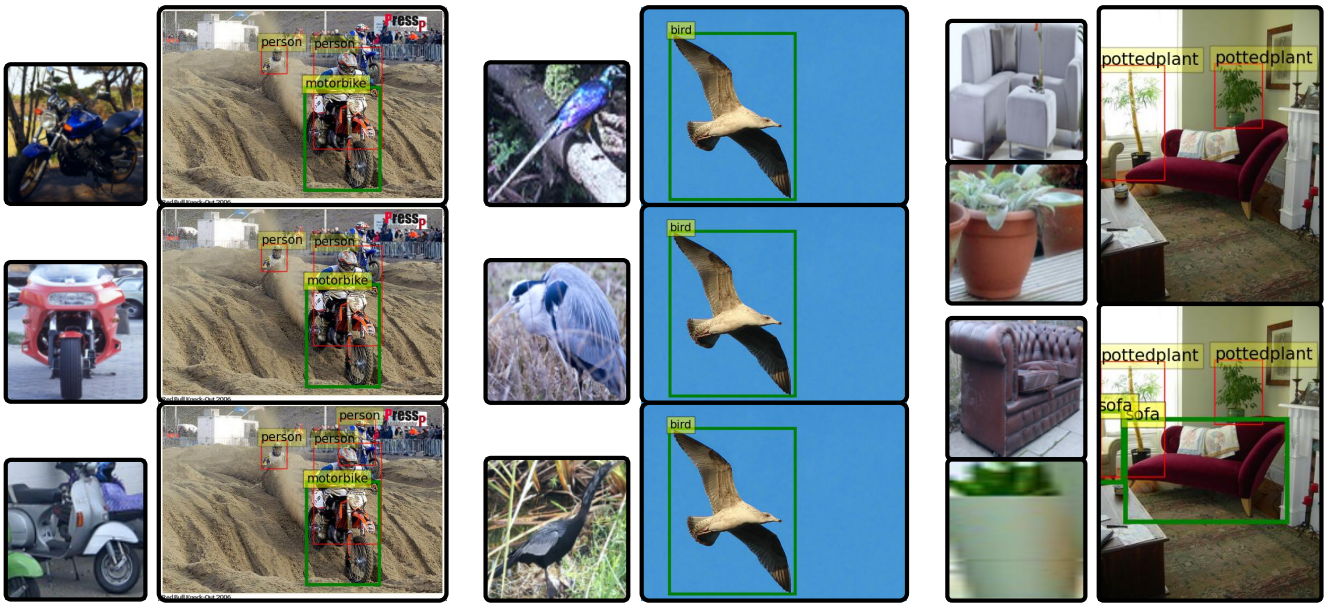

Figure 3. Effect of different 1-shot template on detection with FS-DETR. Small images indicate the template used to detect the objects on the larger images. The left-most two pairs of columns illustrate the robustness to template change, while the right-most column pair illustrates a failure case.

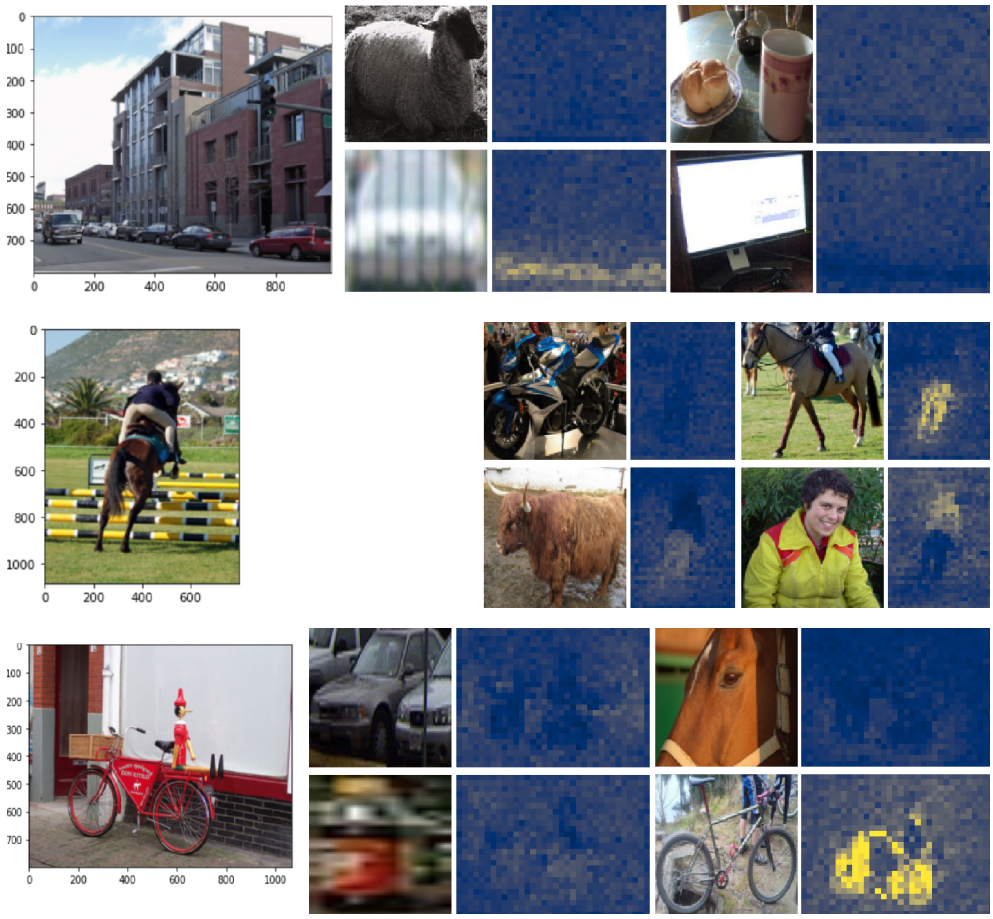

Figure 4. Attention weights between the visual prompts (templates) and the encoded image features for three randomly sampled target images (left column) from VOC Pascal dataset. Notice that the network learns to attend to the parts of the image that are semantically close to the presented templates. For each target image (left column), we show the attention weights generated by four templates. We observe that for the target image of the first row, only the car template generates attention of high magnitude at several locations corresponding to the location of the cars in the target image. Similarly, for the target image of the second row only the horse and the person templates fire at the corresponding locations in the target image as expected. Similar conclusions can be drawn for the target image of the last row.

References

[1] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *CVPR*, pages 14605–14615, 2022. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020. 1

[4] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. UP-DETR: Unsupervised pre-training for object detection with transformers. In *CVPR*, 2021. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[6] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-RPN and multi-relation detector. In *CVPR*, 2020. 1, 2

[7] Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, and Shih-Fu Chang. Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In *ICCV*, 2021. 1, 2

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[11] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional DETR for fast training convergence. In *ICCV*, 2021. 1

[12] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999. 1

[13] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. DeFRCN: Decoupled Faster R-CNN for few-shot object detection. In *ICCV*, 2021. 1, 2

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 2

[16] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 1

[17] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 1