# Cross-Scale Query-Support Alignment Approach for Small Object Detection in the Few-Shot Regime


**Pierre Le Jeune**
*L2TI, Université Sorbonne Paris Nord*
*COSE*

**Anissa Mokraoui**
*L2TI, Université Sorbonne Paris Nord*



30th IEEE International Conference on Image Processing
*Kuala Lumpur – October 8, 2023*

# Overview of the presentation

# 1. Small Objects in Few-Shot Detection

**Small objects are more difficult to detect.**
Experimental evidences reported in many detection frameworks, e.g. Faster R-CNN (Ren et al. 2015), YOLO (Redmon et al. 2016) or DETR (Carion et al. 2020).

**Difficulty greatly reinforced in Few-Shot Regime (Le Jeune and Mokraoui 2022).**

- Small objects are poor examples for the model and miscondition the detection.
- Greater performance gap between small and large objects in Few-Shot Regime.
- Explains the performance gap between natural and aerial images in FSOD.

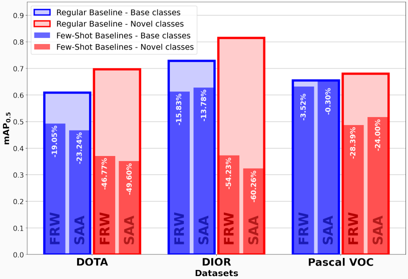


**Figure 1:** Few-Shot Detection performance compared on three distinct datasets, figure from (Le Jeune and Mokraoui 2022).

# 2. Attention-based Few-Shot Object Detection Principle

**Few-Shot Object Detection literature is mainly based on attention mechanism.**

**Key principle:**

Adapt features from the query image on-the-fly during inference from few annotated support examples.

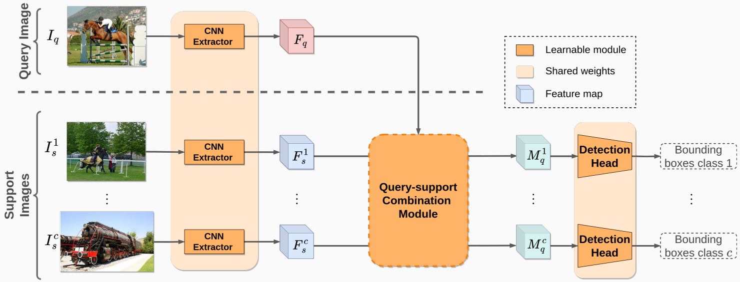


**Figure 2:** Attention-based FSOD principle.

# 2. Attention-based Few-Shot Object Detection Principle

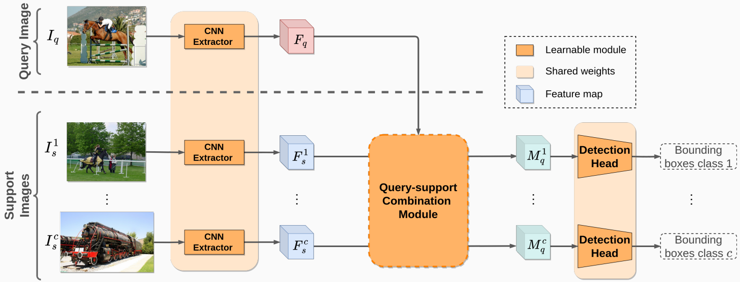


**Figure 2:** Attention-based FSOD principle.

Three main components:

- **Backbone:** extracts features from the image.
- **Query-Support Combination Module:** combines query and support features.
- **Detection Head:** performs object detection in a class-agnostic manner.

# 3.1 Cross-scale Query-Support Alignment (XQSA) - Overall Structure

**Objective:** propose a better Query-Support combination block to improve small object detection.

Query-Support combination modules are often split into three components:

- **Self Attention:** combines query and support features at a global scale.
- **Spatial Alignment:** locally compares features from query and support.
- **Feature Fusion:** aggregates relevant information for detection.

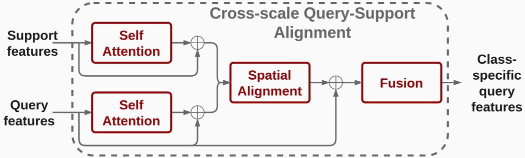


**Figure 3:** Overall structure of the Cross-Scale Query-Support Alignment block.

# 3.1 Cross-scale Query-Support Alignment (XQSA) - Overall Structure

**Objective:** propose a better Query-Support combination block to improve small object detection.

Query-Support combination modules are often split into three components:

- **Self Attention:** combines query and support features at a global scale.
- **Spatial Alignment:** locally compares features from query and support.
- **Feature Fusion:** aggregates relevant information for detection.

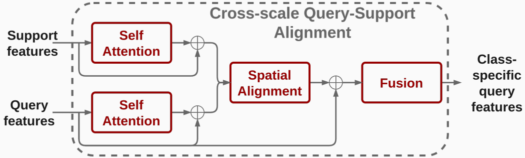


**Figure 3:** Overall structure of the Cross-Scale Query-Support Alignment block.

## XQSA's motivation

- Combines query and support features from different scales together.
- Allows matching query and support objects from different sizes.

$\rightarrow$ *properties not available in the literature.*

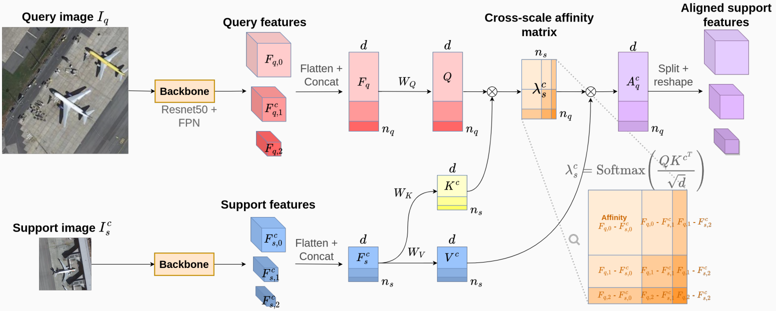


**Figure 4:** Illustration of the Spatial Alignment block in XQSA.

### Global Attention (self-attention)

$$\psi = \text{SA}(\phi) = \text{Softmax}(\phi W_{\text{SA}}) \cdot \phi$$

### Spatial Alignment

$$Q = F_q W_Q = [F_{q,0}, F_{q,1}, F_{q,2}] W_Q,$$
$$K^c = F^c_s W_K = [F^c_{s,0}, F^c_{s,1}, F^c_{s,2}] W_K,$$
$$V^c = F^c_s W_V = [F^c_{s,0}, F^c_{s,1}, F^c_{s,2}] W_V.$$

$$\lambda^c = \text{Softmax}\left(\frac{QK^{cT}}{\sqrt{d}}\right),$$

$$A^c_q = \lambda^c V^c.$$

### Feature Fusion

$$M^c_{q,l} = [F_{q,l}, A^c_{q,l}] W^l_F$$

## 4.1 Few-Shot Performance on Natural and Aerial Images

Comparison with two existing methods: **FRW** (Kang et al. 2019) and **DANA** (Chen et al. 2021).
Two aerial datasets **DOTA** (Xia et al. 2018) and **DIOR** (Li et al. 2020), and two natural datasets **Pascal VOC** (Everingham et al. 2010) and **MS COCO** (Lin et al. 2014)

| | | DOTA | | | | DIOR | | | | Pascal VOC | | | | MS COCO | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All | S | M | L | All | S | M | L | All | S | M | L | All | S | M | L |
| Base Classes | FRW | 49.04 | 25.48 | 59.17 | 63.37 | 62.20 | 8.21 | 48.66 | 80.67 | 63.21 | 15.67 | 47.94 | **81.73** | 29.03 | 13.08 | 35.87 | 48.00 |
| | DANA | **53.98** | **37.00** | **62.27** | **70.32** | **62.71** | **10.92** | **49.34** | **83.17** | **65.17** | **18.14** | **50.58** | 80.11 | **38.14** | **23.30** | **51.85** | **56.38** |
| | XQSA | 51.11 | 26.10 | 59.41 | 64.30 | 59.88 | 10.64 | 45.69 | 82.34 | 62.13 | 15.60 | 48.64 | 75.94 | 31.56 | 16.13 | 40.13 | 49.83 |
| Novel Classes | FRW | 37.29 | 13.99 | 34.11 | 59.31 | 36.29 | 2.48 | 33.74 | 59.38 | 48.72 | 16.44 | 26.71 | **68.27** | 24.09 | 11.53 | 22.45 | **38.69** |
| | DANA | 36.38 | 14.33 | 40.00 | **64.64** | 38.18 | 3.21 | 34.91 | **60.99** | 52.26 | 10.05 | 24.67 | 67.23 | 24.75 | 12.01 | **29.40** | 37.95 |
| | XQSA | **41.00** | **17.84** | **44.57** | 54.46 | **41.51** | **4.12** | **40.69** | 58.21 | **53.94** | **19.46** | **34.86** | 66.14 | **25.03** | **12.57** | 26.05 | 38.55 |

**Table 1:** Performance comparison between XQSA, FRW, and DANA. $mAP_{0.5}$ values are reported separately for base (top) and novel (bottom) classes on DOTA, DIOR, Pascal VOC, and MS COCO with $K = 10$ shots. mAP values are reported for All, Small ($\sqrt{wh} < 32$), Medium ($32 \leq \sqrt{wh} < 96$) and Large ($\sqrt{wh} \geq 96$) objects.

**$\rightarrow$ XQSA largely improves the detection of small objects both for natural and aerial images in the few-shot regime.**

$\rightarrow$ Improvement at the cost of performance on larger objects, but overall it helps a lot for aerial images.

# 4.2 Ablation Study

Each component of the Cross-Scale Query-Support Alignment (XQSA) block is useful to improve novel classes detection performance.

Performance on base classes are also improved to a lesser extent.

| | | | | |
|---|---|---|---|---|
| Baseline | ✓ | ✓ | ✓ | ✓ |
| Cross-scale Alignment | | ✓ | ✓ | ✓ |
| Fusion Layer | | | ✓ | ✓ |
| Query-Support Self-Attention | | | | ✓ |
| **Base classes** | 49.20 | 49.46 | 49.13 | **51.11** |
| **Novel classes** | 36.52 | 38.84 | 40.31 | **41.01** |

**Table 2:** Ablation study of the XQSA attention method on DOTA dataset. $\text{mAP}_{0.5}$ scores are reported for base and novel classes with $K = 10$ shots.

# 5. Conclusion and perspectives

### Key takeaways

- XQSA largely improves the detection performance of small object in the few-shot regime.
- Improvement on small target at the cost of larger objects.
- Very helpful for aerial images.

### Perspectives

- Design attention mechanism for both small and large objects.
- While XQSA improves on small objects, the performance on small target is still low, adjustments required in other part of the framework (e.g. backbone, detection head, or training procedure).

## Thank you for your attention

Any questions ?

pierre.lejeune@edu.univ-paris13.fr

https://pierlj.github.io

# References i


Carion, Nicolas et al. (2020). “End-to-end object detection with transformers”. In: *European Conference on Computer Vision*. Springer, pp. 213–229.

Chen, Tung-I et al. (2021). “Should I Look at the Head or the Tail? Dual-awareness Attention for Few-Shot Object Detection”. In: *arXiv preprint arXiv:2102.12152*.

Everingham, Mark et al. (2010). “The pascal visual object classes (voc) challenge”. In: *International journal of computer vision* 88.2, pp. 303–338.

Kang, Bingyi et al. (2019). “Few-shot object detection via feature reweighting”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8420–8429.

Le Jeune, Pierre and Anissa Mokraoui (2022). “Improving Few-Shot Object Detection through a Performance Analysis on Aerial and Natural Images”. In: *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*.

Li, Ke et al. (2020). “Object detection in optical remote sensing images: A survey and a new benchmark”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 159, pp. 296–307.

Lin, Tsung-Yi et al. (2014). “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer, pp. 740–755.

Redmon, Joseph et al. (2016). “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

Ren, Shaoqing et al. (2015). “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28, pp. 91–99.

Xia, Gui-Song et al. (2018). “DOTA: A large-scale dataset for object detection in aerial images”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3974–3983.