Few Shot Learning for Vision

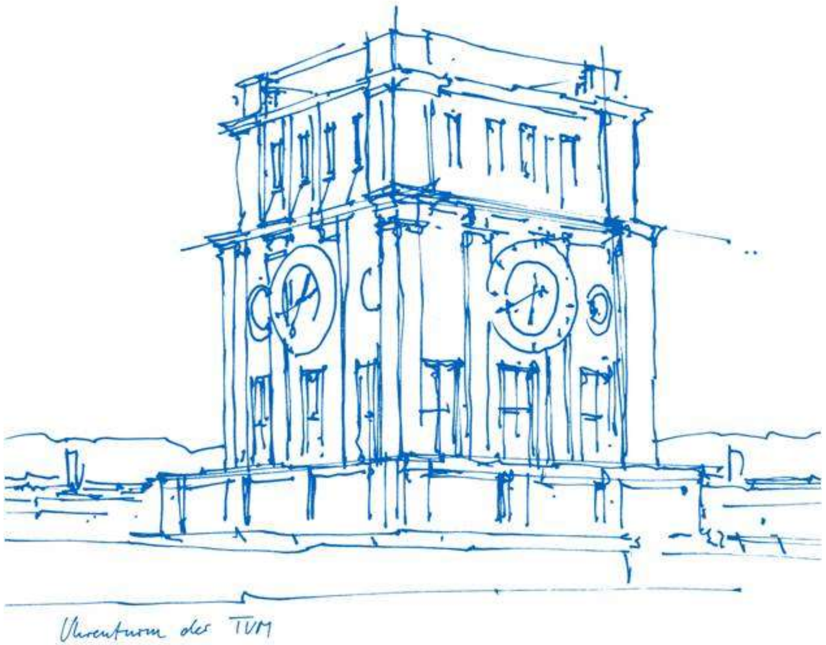
Alaa Arbi

Technical University Munich

TUM School of Computation, Information and Technology

Chair of Computer Visio & Artificial Intelligence

Munich, 17. and 18. January 2023



Outline

1. Motivation
2. Problem Definition
3. Training Paradigms
4. Localization
5. Classification
6. Datasets
7. Results

References

Motivation

- Deep learning based detectors accurate
- Large-scale datasets
- Costly or impossible
- Humans able to learn from few samples
- Few Shot Object Detection

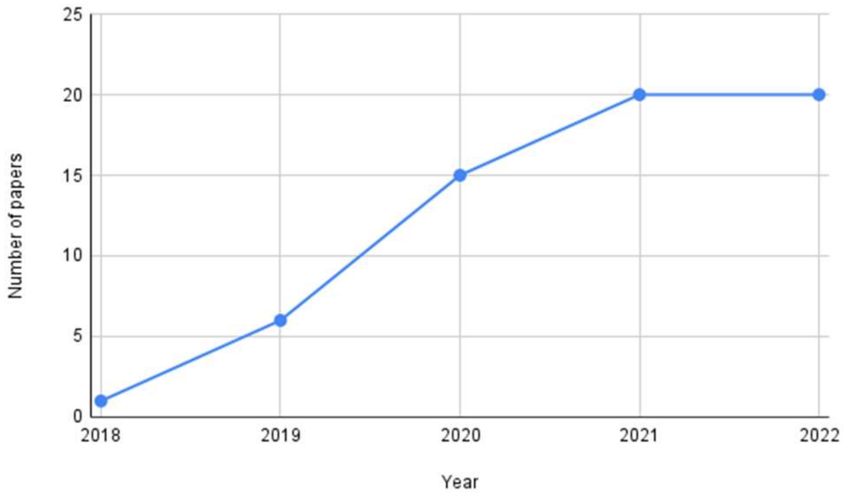


Figure 1: Number of few-shot object detection papers covered in survey [4] by year of publication

Problem Definition

- D_{base}: large dataset (COCO, Pascal Voc etc.)
- D_{novel}: Dataset with only K object instances per category
- $C_{base} \cap C_{novel}$ = None
- D_{query}: query dataset
  - FSOD: $C_{query} = C_{novel}$
  - G-FSOD: $C_{query} = C_{base} \cup C_{novel}$

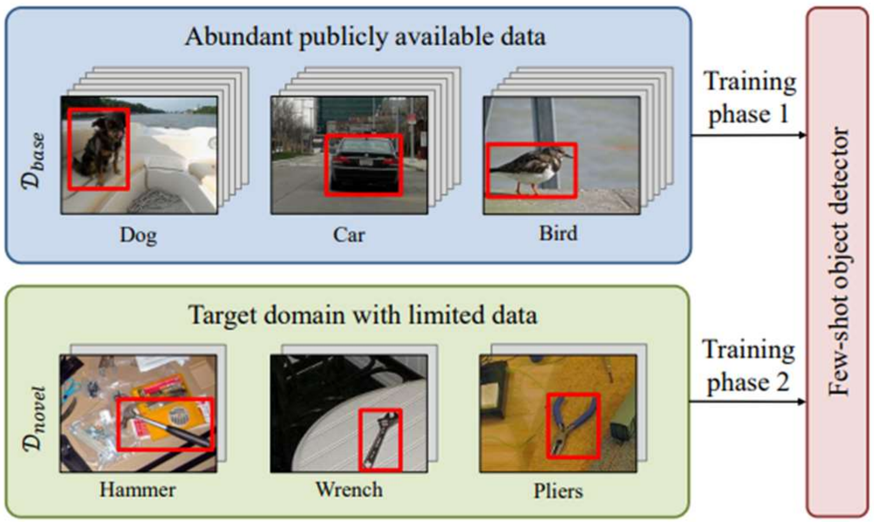


Figure 2: General setting of few shot object detection: First train on a large dataset then fine-tune on the small dataset, Image taken from [4]

Training Paradigms

Two-Stage training:

1. Train M_{init} using D_{base} M_{base}
2. Fine-tune M_{base} using $D_{finetune}$ $M_{finetune}$

Meta-Learning:

E episodes

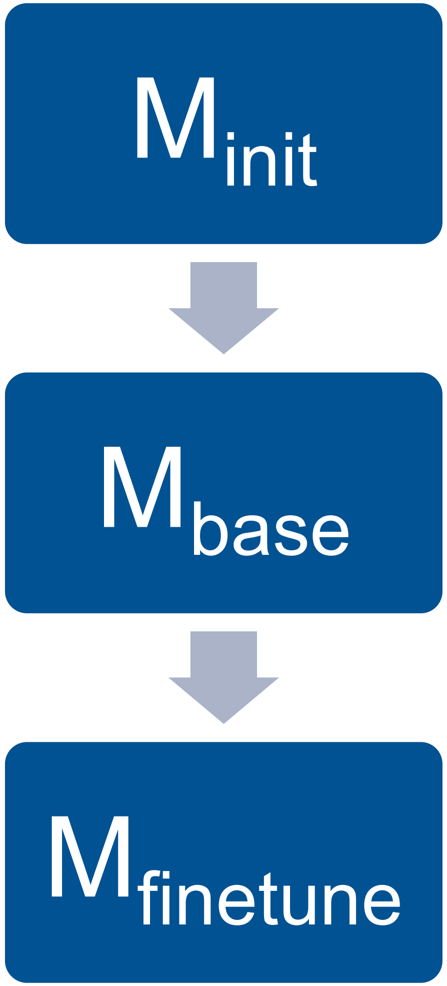
K instances randomly sampled

Transfer-Learning:

Base training

Freezing

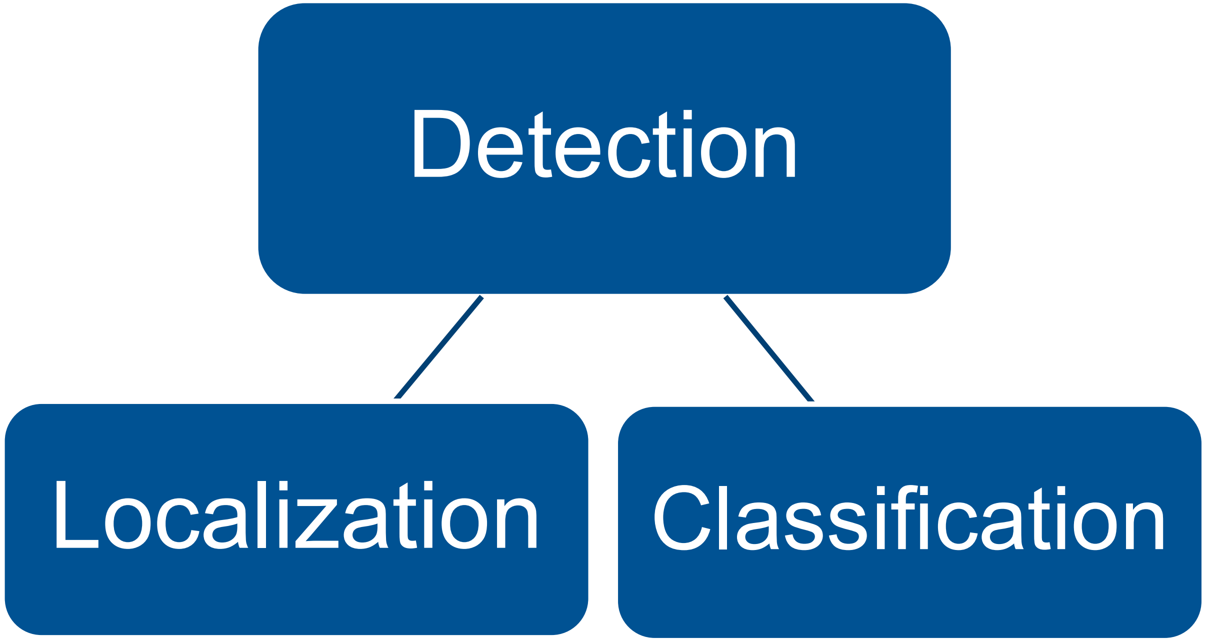
Finetuning



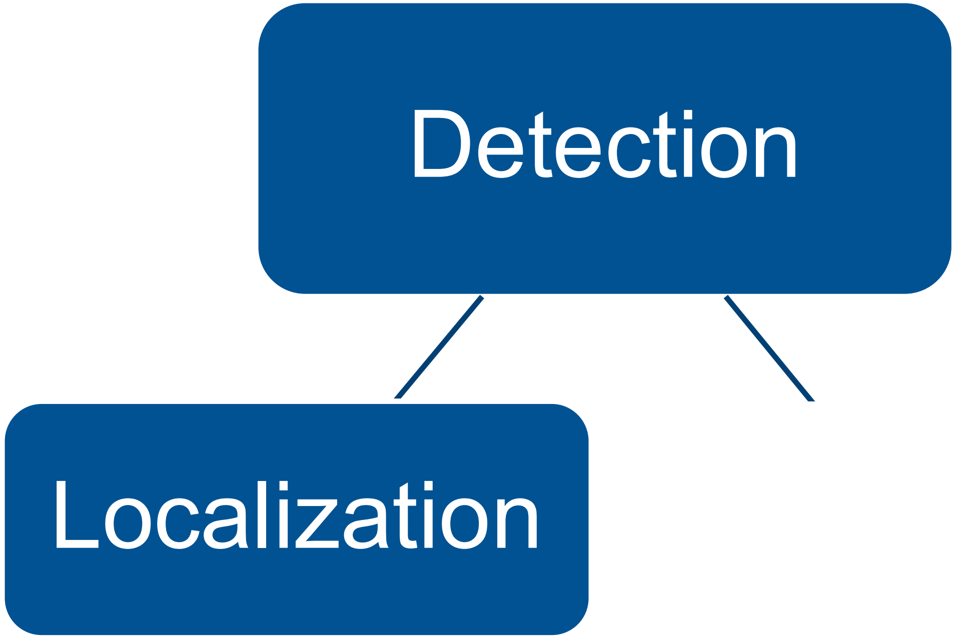
Training Paradigms

DeFRCN	**MetaDETR**
Based on the Faster R-CNN Trained using Transfer-Learning	Based on the DETR (based on transformers) Trained using Meta-Learning

Object Detection



Object Detection



Localization

Localization (C_{novel}) < Localization (C_{base})

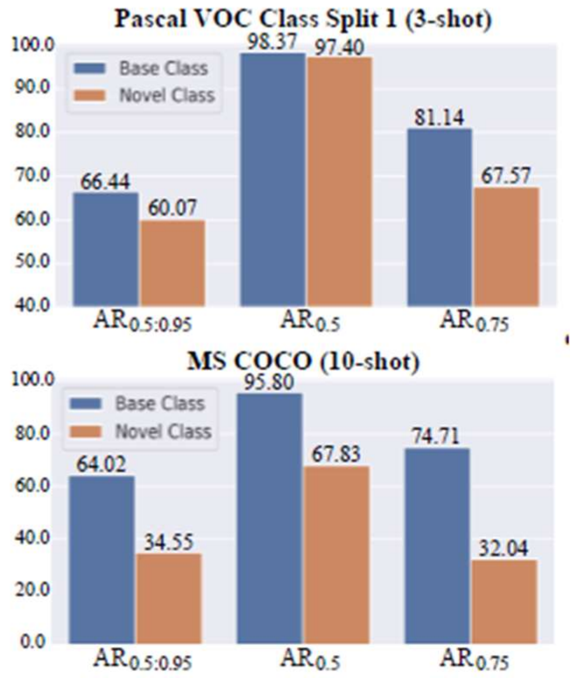


Figure 3: Average Recall on the top 1000 region proposals for novel and base classes Image taken from [1]

Localization

DeFRCN

Contradiction in network components:

- RPN: translation co-variant
- Classification head: translation invariant

Foreground-Background confusion in PR:

- Novel classes = background in base training

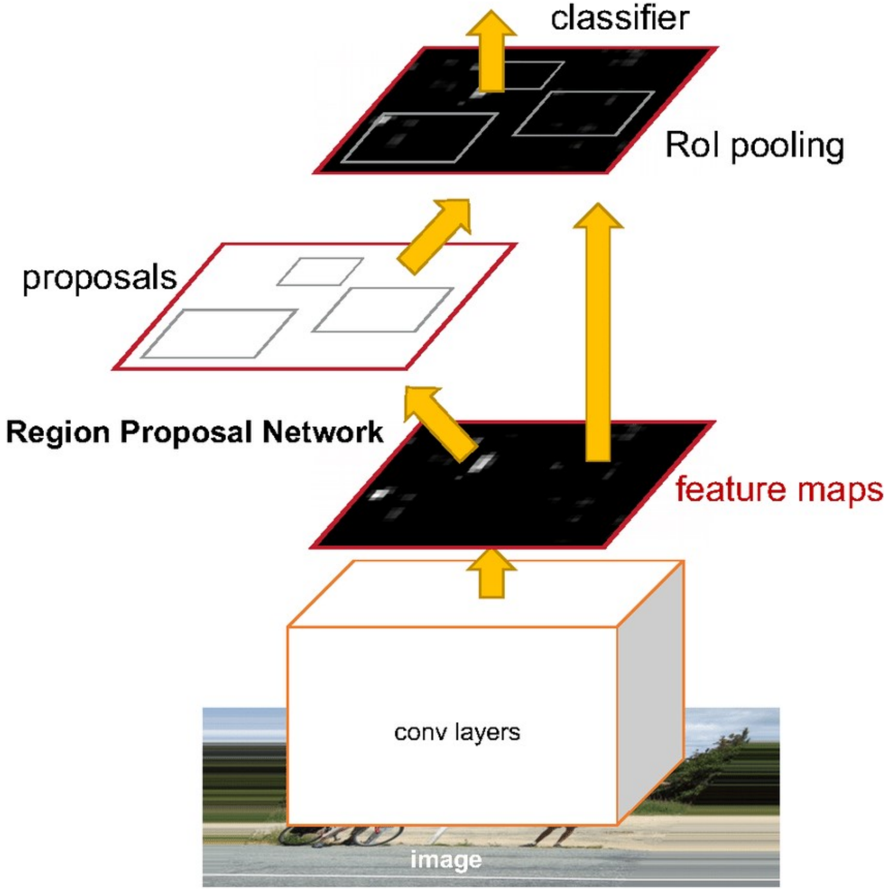


Figure 4: Architecture of Faster R-CNN for general object detection. Image taken from [5]

Localization

DeFRCN

Use Gradient Decoupled Layer to:

1. specific features for RPN and RCNN
2. Stop gradient flow from RPN to backbone
3. Scale gradient flow RCNN to backbone

Result: No domination by one part

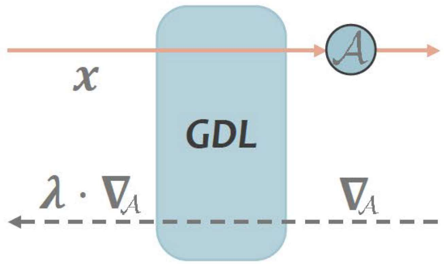


Figure 6: Gradient Decoupled Layer. Image taken from [2]

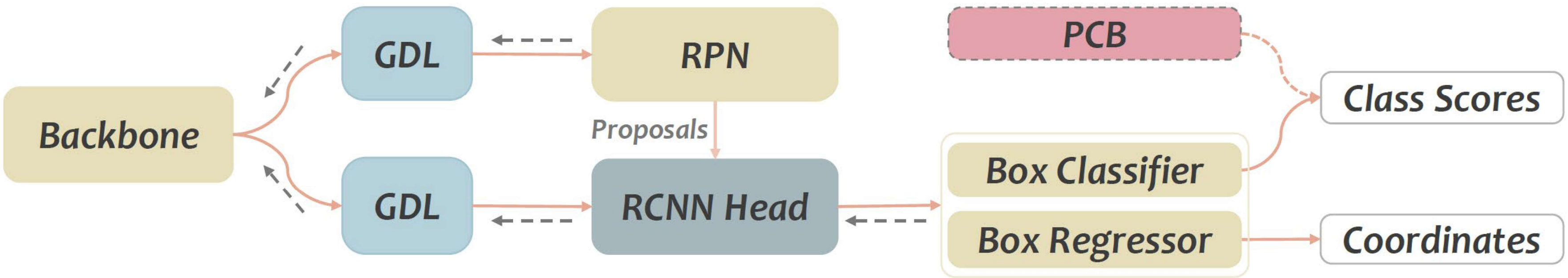


Figure 5: Architecture of DeFRCN. Image taken from [2]

Localization

MetaDETR

No Region Proposals
Image Level Detector
DETR architecture: based on attention

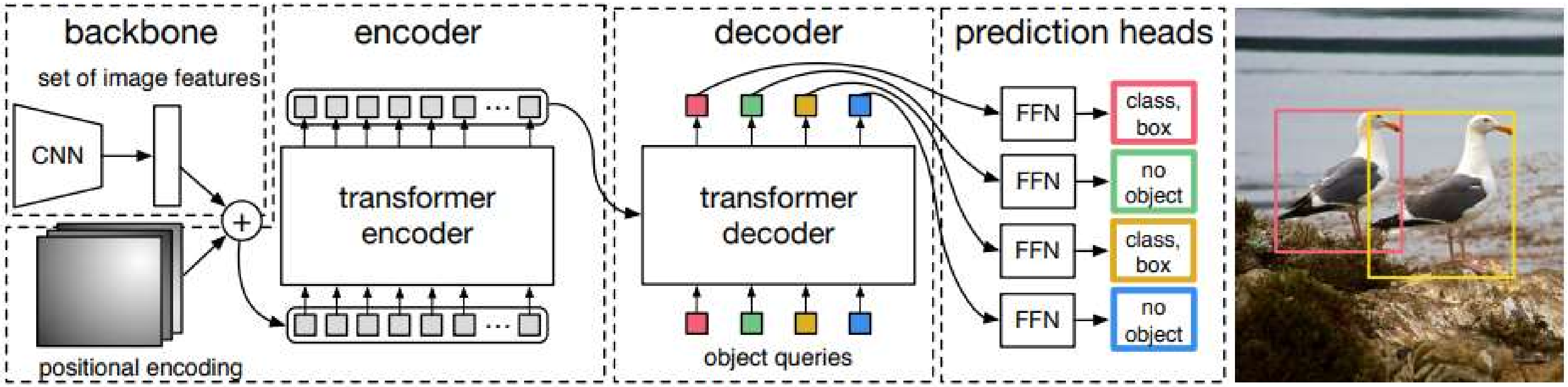


Figure 7: The DETR architecture for general object detection. Image taken from [3]

Localization

MetaDETR

Self-Attention:
Inter-Relation of Image features

Cross Attention:
Match Image features with object queries

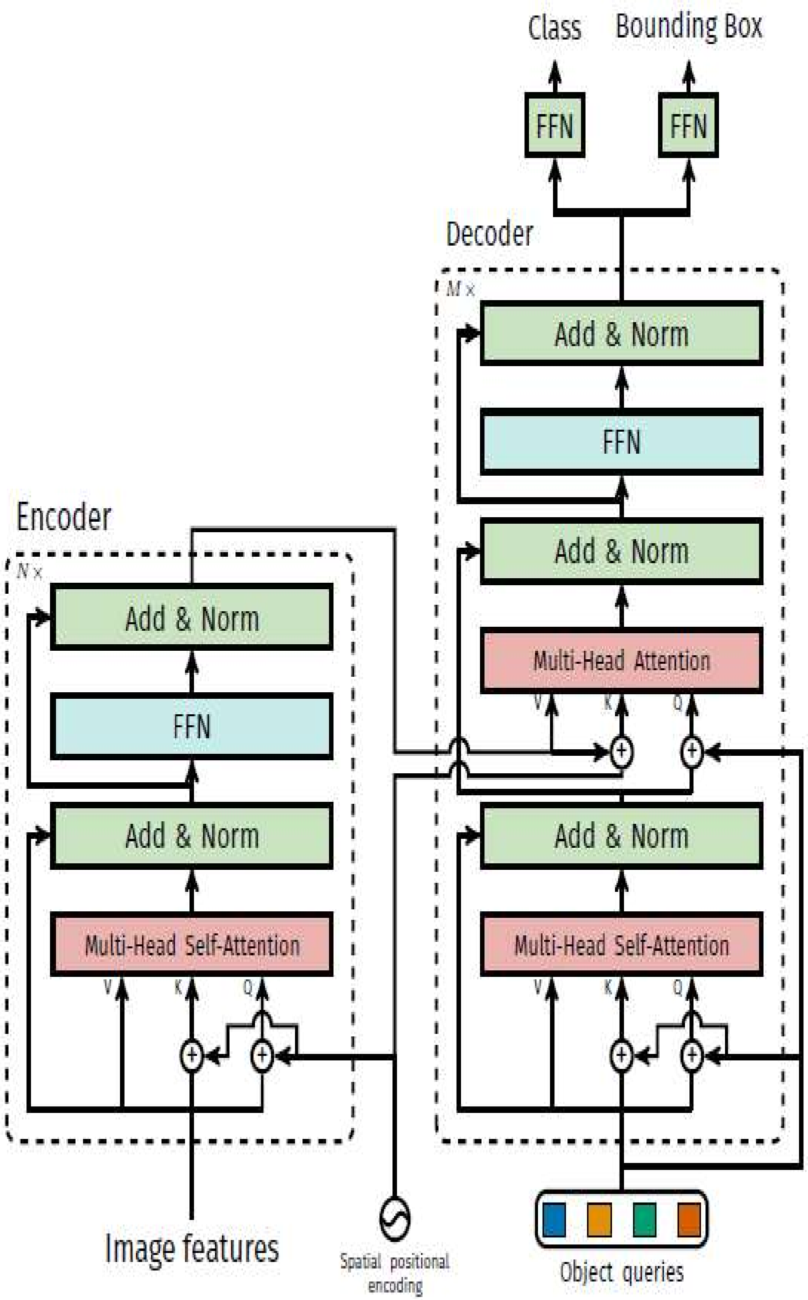


Figure 8: Details of the Encoder-Decoder transformer used in DETR. Image taken from [3]

Localization

MetaDETR

Self-Attention:
Inter-Relation of Image features

Cross Attention:
Match Image features with object queries

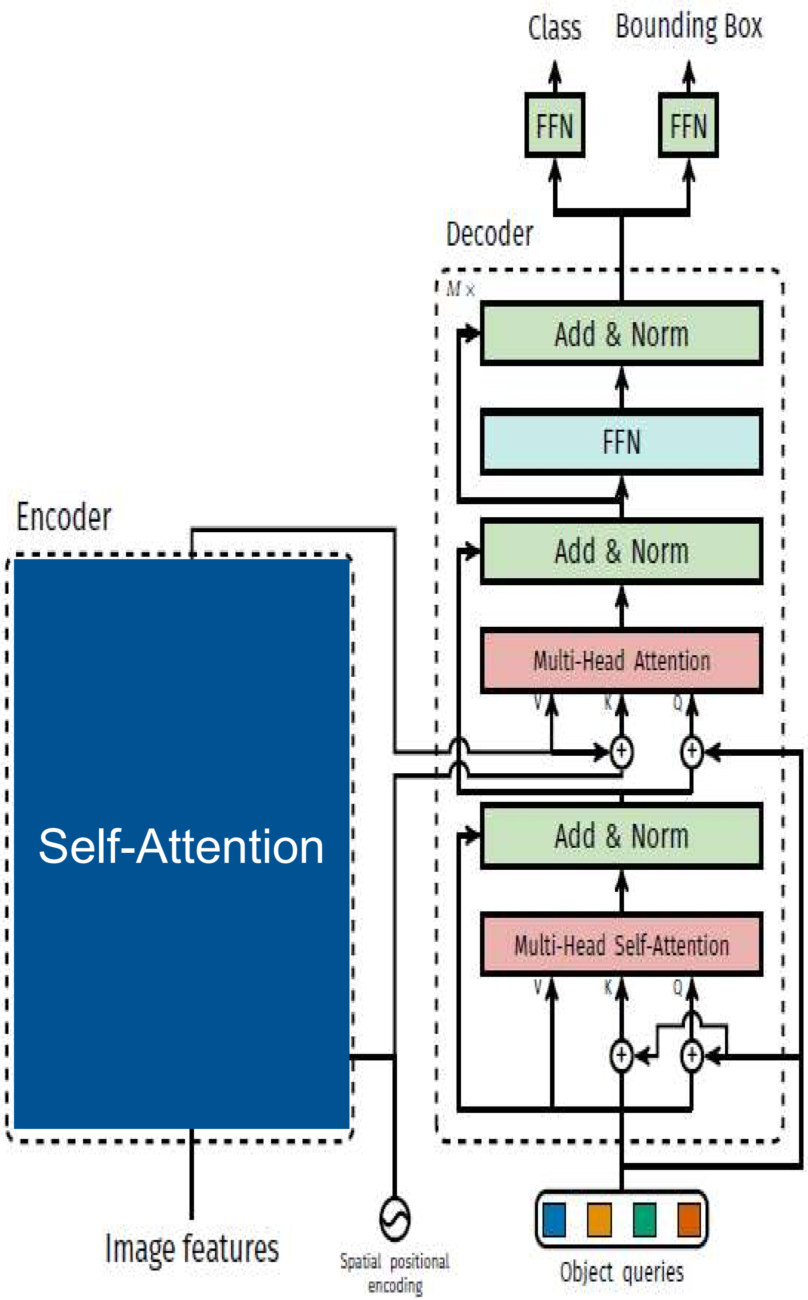


Figure 8: Details of the Encoder-Decoder transformer used in DETR. Image taken from [3]

Localization

MetaDETR

Self-Attention:
Inter-Relation of Image features

Cross Attention:
Match Image features with object queries

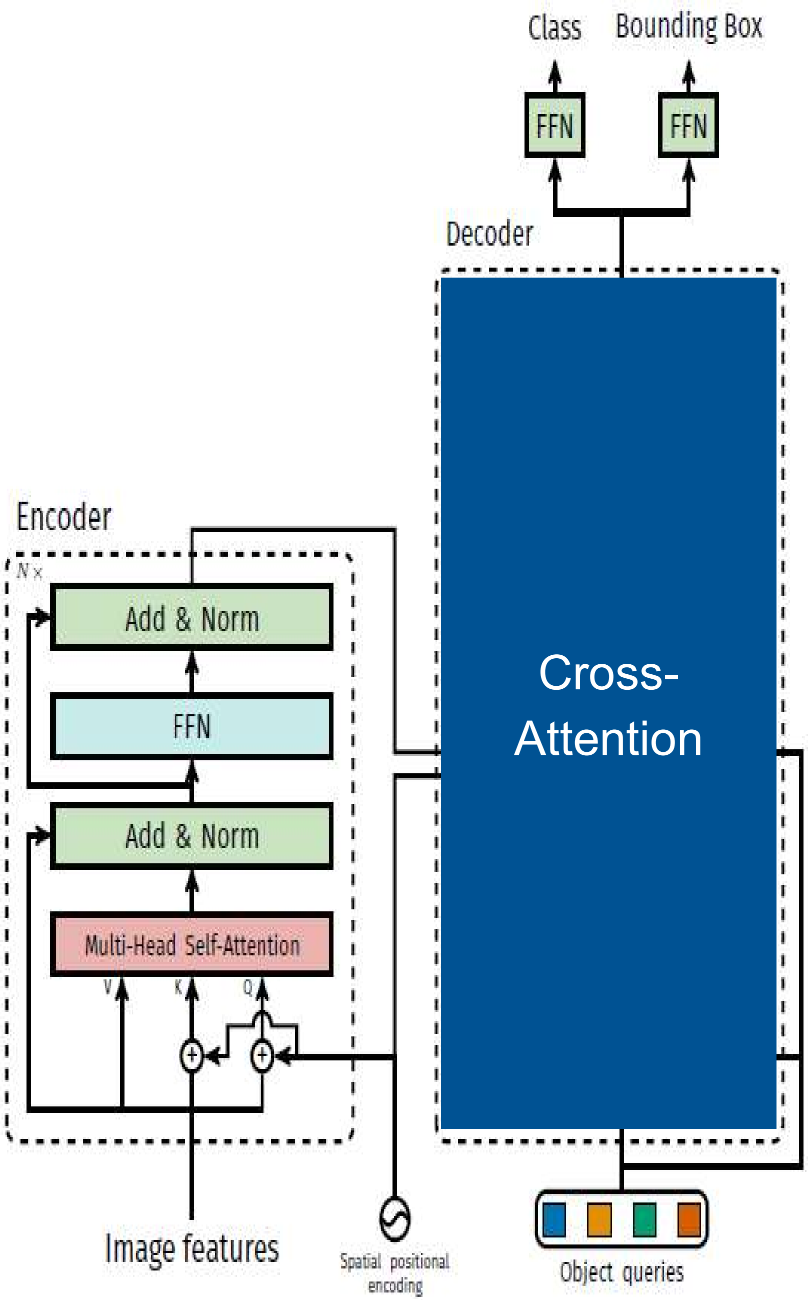


Figure 8: Details of the Encoder-Decoder transformer used in DETR. Image taken from [3]

Localization

MetaDETR

The encoder seems to assign high attention coeffients to pixels corresponding to same object and lower ones to other pixels

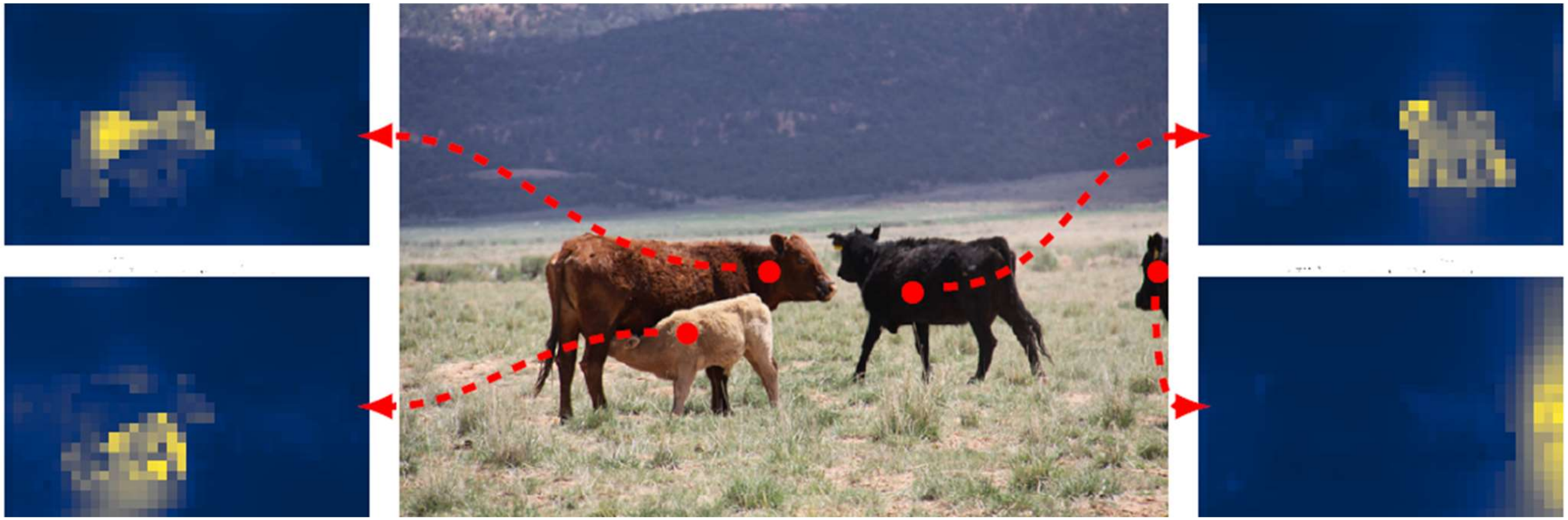


Figure 9: Visualization of the attention map of the encoder. Yellow/Blue indicate a high/low attention value.Image taken from [3]

Localization

MetaDETR

The decoder is assigning high attention coefficients for pixels defining object extremities

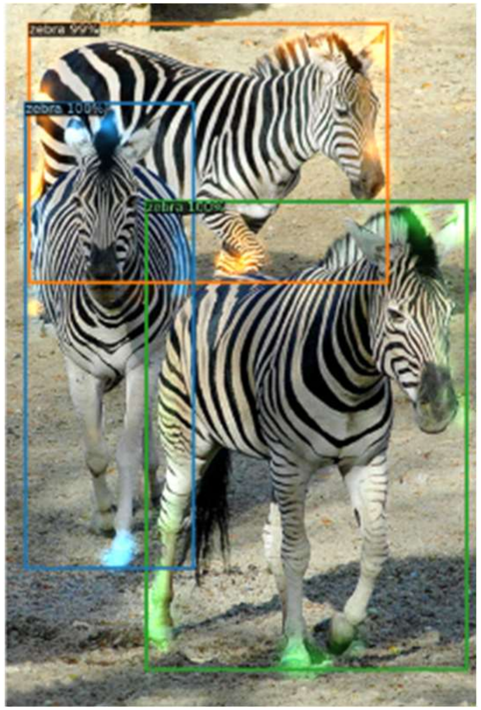
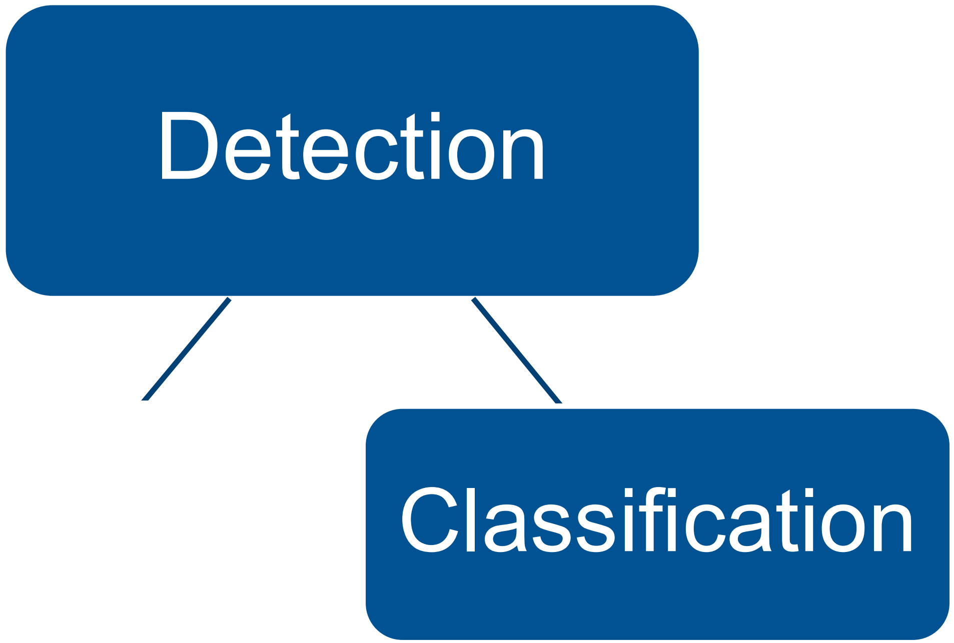


Figure 10: Visualization of the attention map of the decoder.Image taken from [3]

Object Detection



Classification

DeFRCN:
classification scores = low-quality.

MetaDETR:
high missclassfication rates, similar appearances

Figure 11: Missclassified objects because of high appearance similarity. Image taken from [1]

Classification

DeFRCN

Use Prototypical Calibration Block to refine classification scores

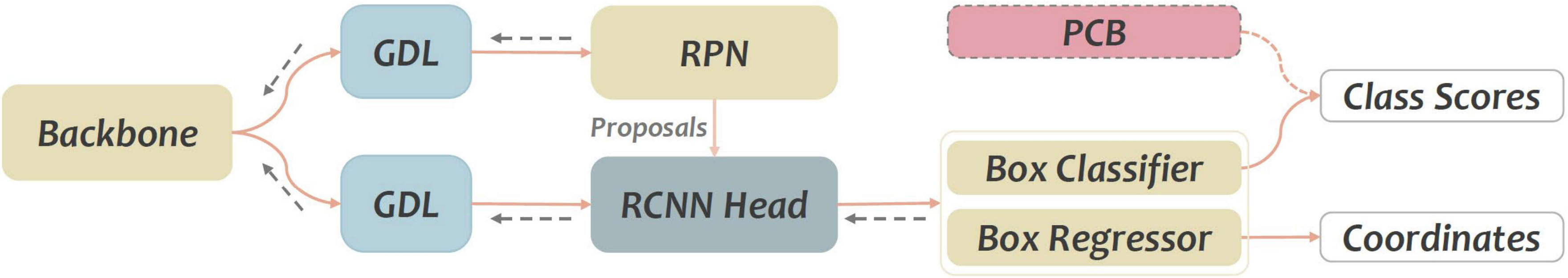


Figure 12: Architecture of DeFRCN. Image taken from [2]

Classification

DeFRCN

The Prototypical Calibration Block:

1. Support set -> class prototypes
2. RoI features from query image
3. Cosine similarity scores
4. Refine classification scores: $s^{refined} = \alpha\, s + (1-\alpha)\, s^{cosine}\ with\ \alpha = 0{,}5$

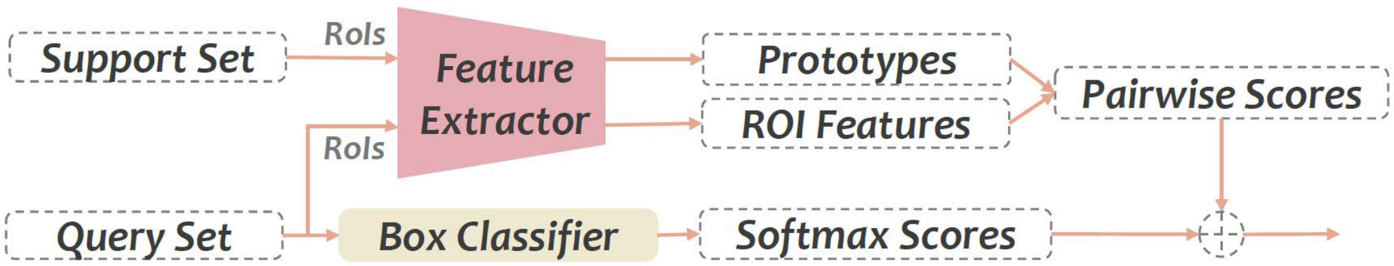


Figure 13: Details of the Prototypical Calibration Block. Image taken from [2]

Classification

DeFRCN

Cosine similarity between class prototypes and features corresponding to specific pixels

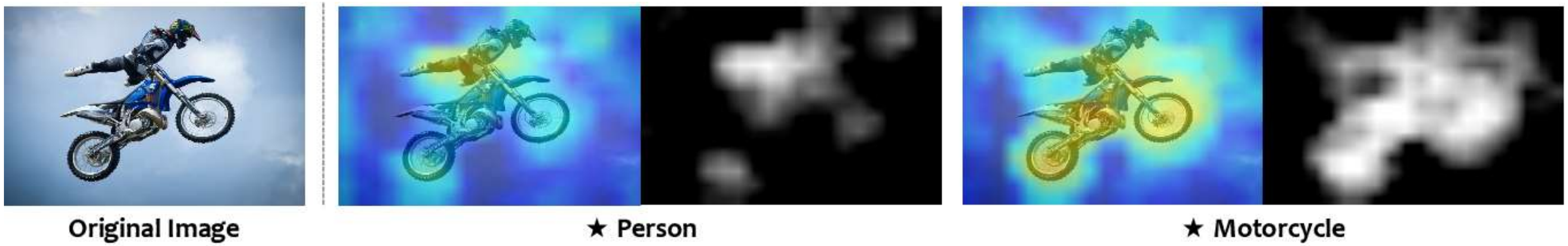


Figure 14: Visualization of the cosine similarity between class prototypes and image features. The white colour indicates a high similarity. Image taken from [2]

Classification

MetaDETR

Use a Correlational Aggregation Module CAM to integrate query features with inter-class correlation information from support images

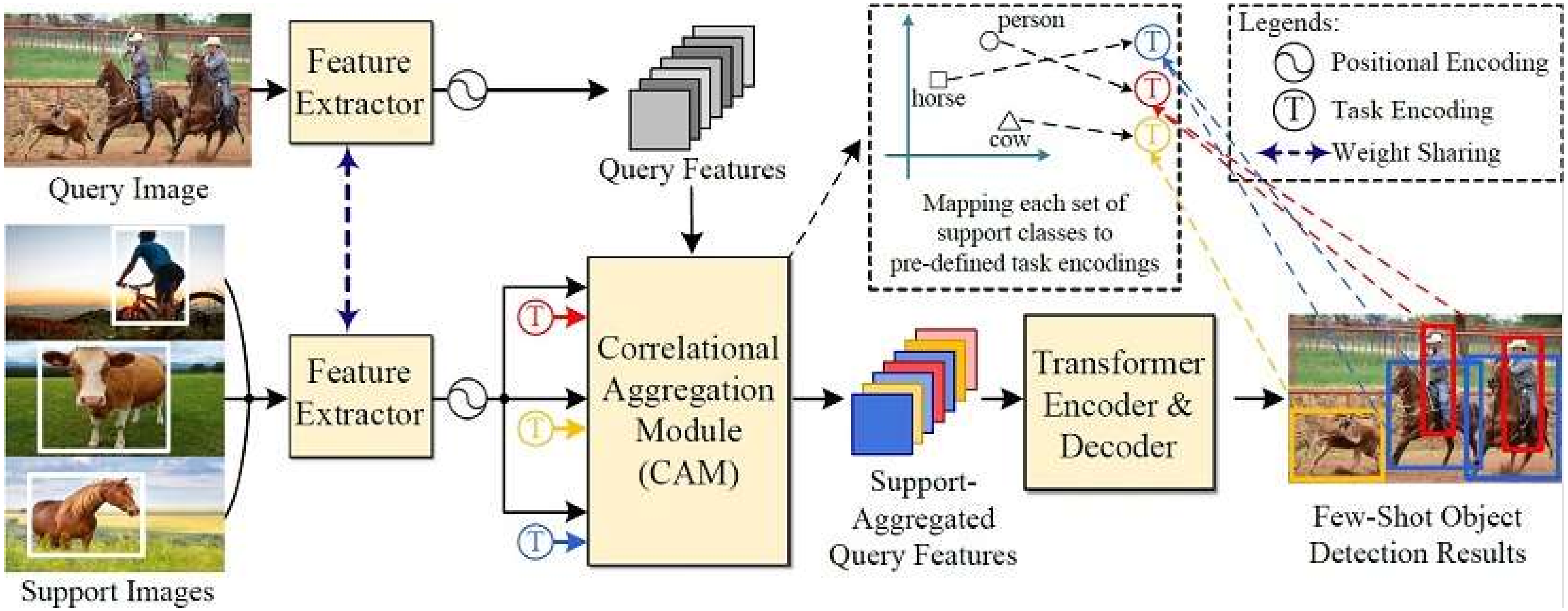


Figure 15: Framework of the MetaDETR. Note how a Correlational Aggregation Module is added between the feature extractor and the Transforer Encoder-Decoder. Image taken from [1]

Classification

MetaDETR

1. Shared Mutli-Head Self-Attention: for Image Query and Support Class features
2. Feature Matching
3. Encoding Matching
4. Merge both featues using elementwise addition

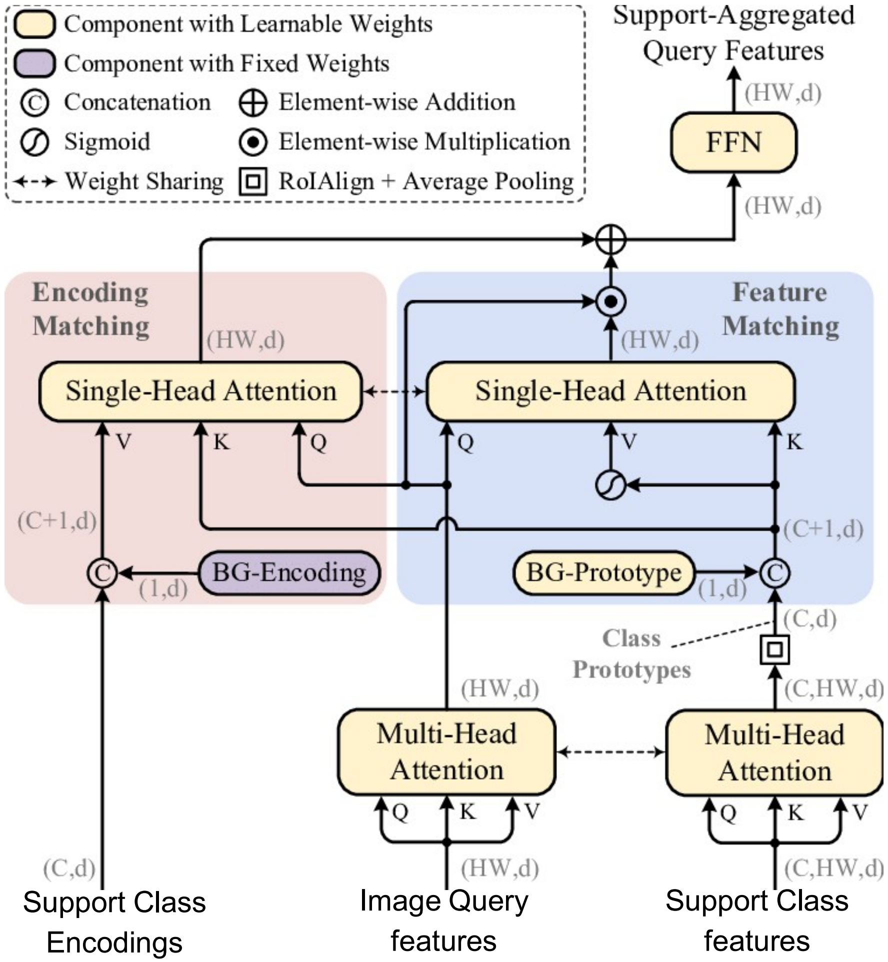
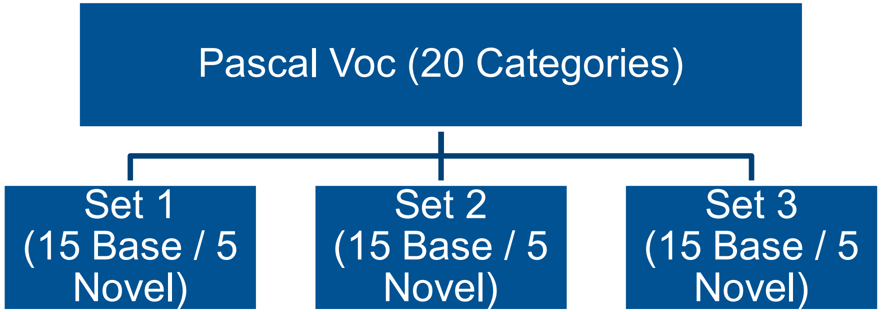
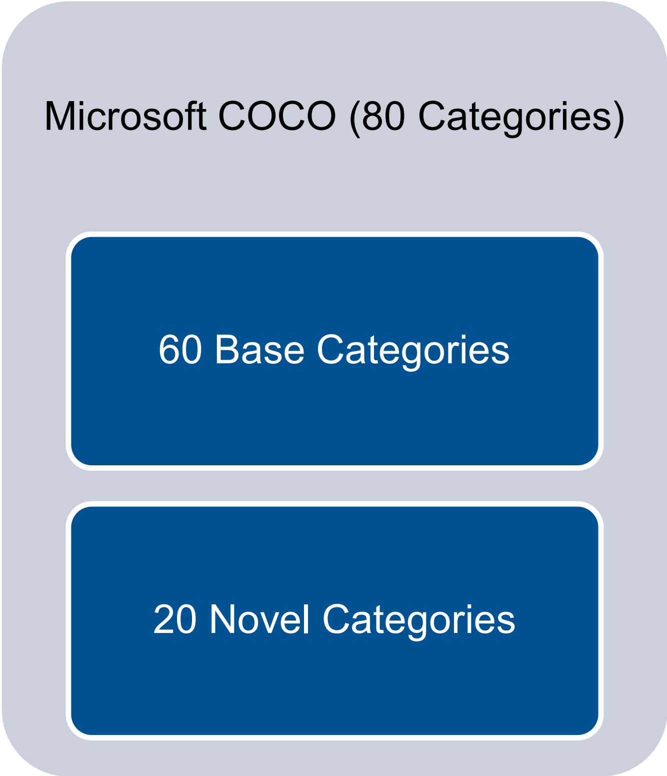


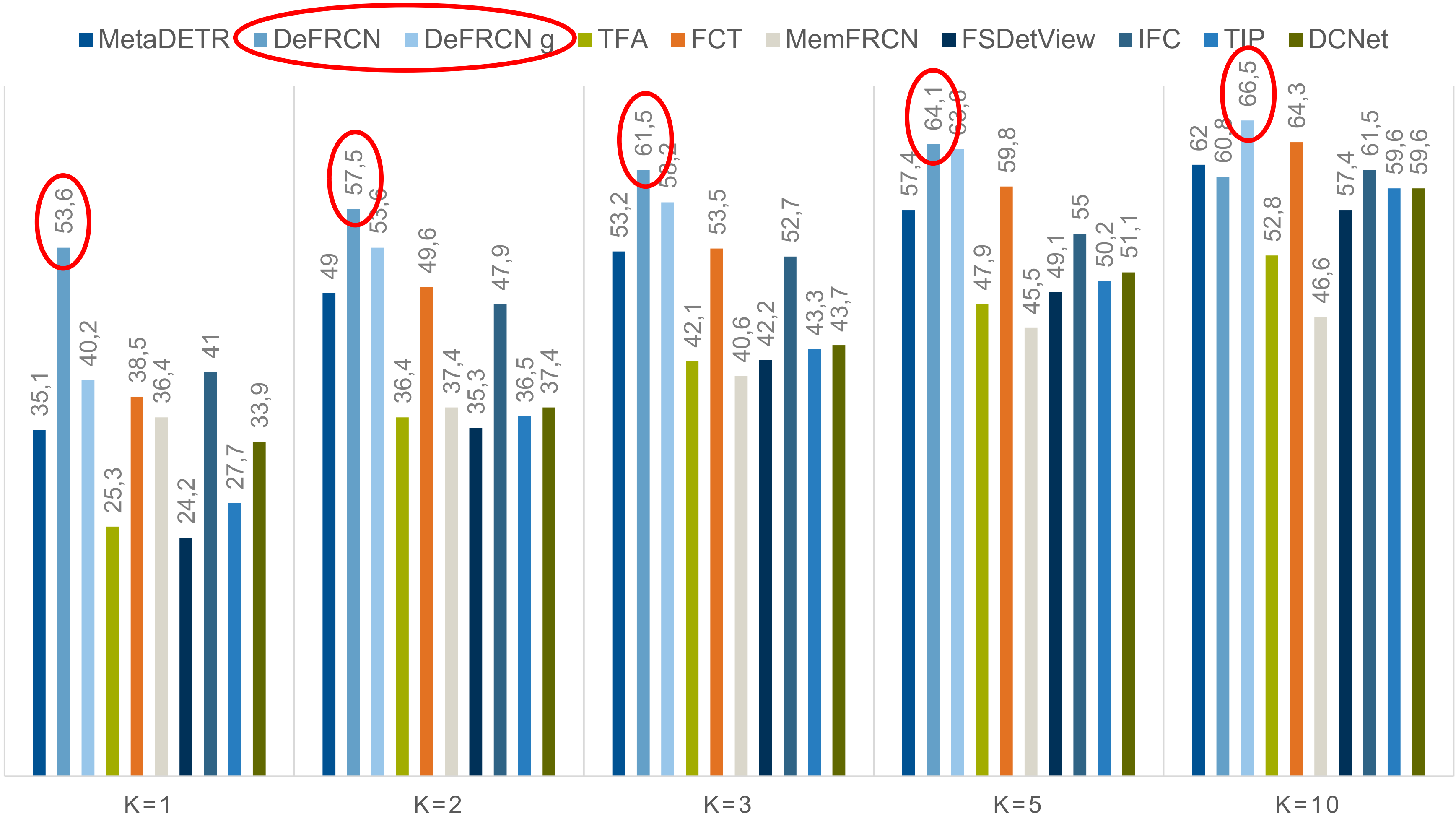
Figure 16: Details of Correlational Aggregation Module
.Image taken from [1]

Datasets

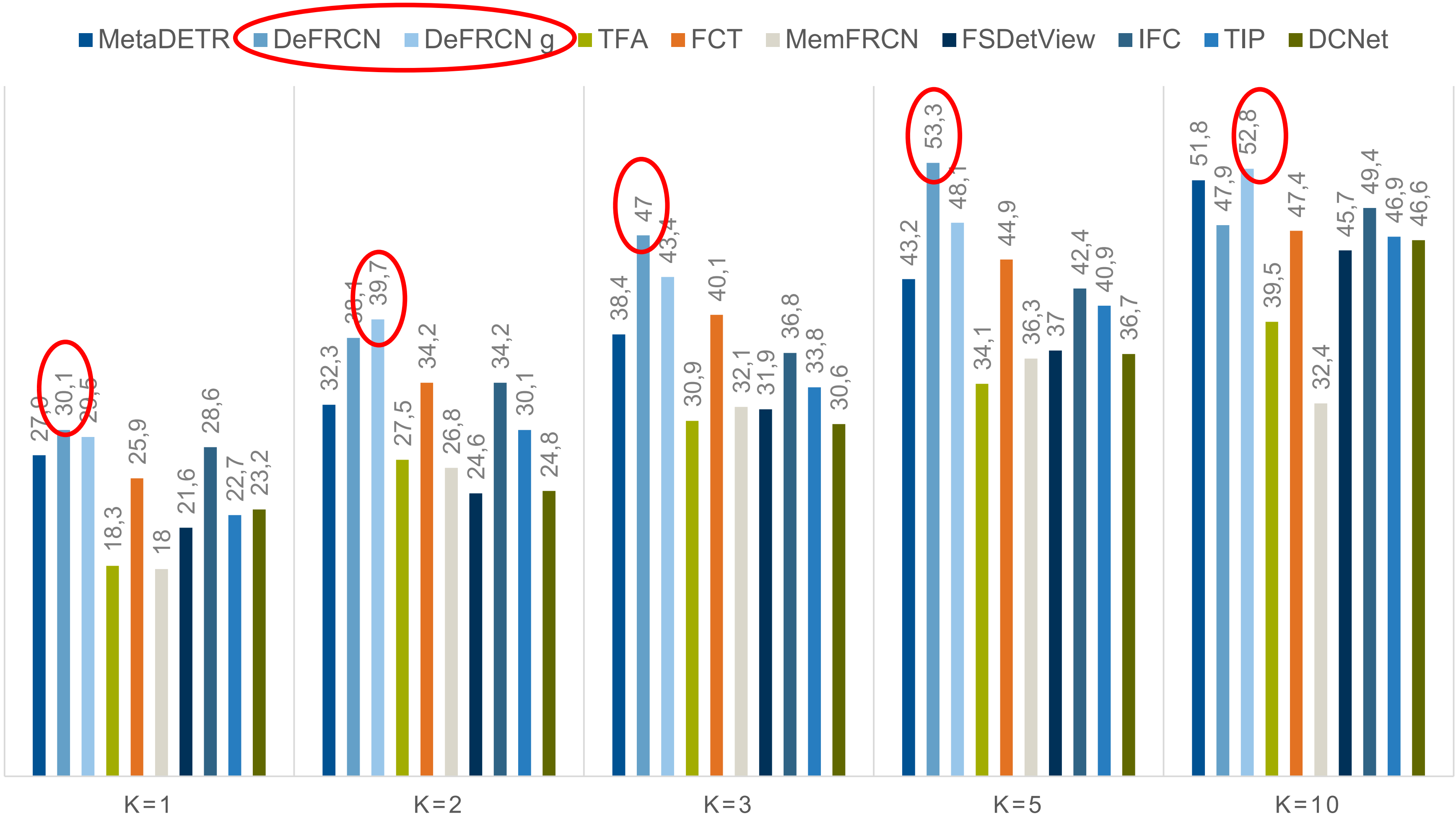

Pascal Voc (20 Categories)
Set 1 (15 Base / 5 Novel)
Set 2 (15 Base / 5 Novel)
Set 3 (15 Base / 5 Novel)


Microsoft COCO (80 Categories)
60 Base Categories
20 Novel Categories

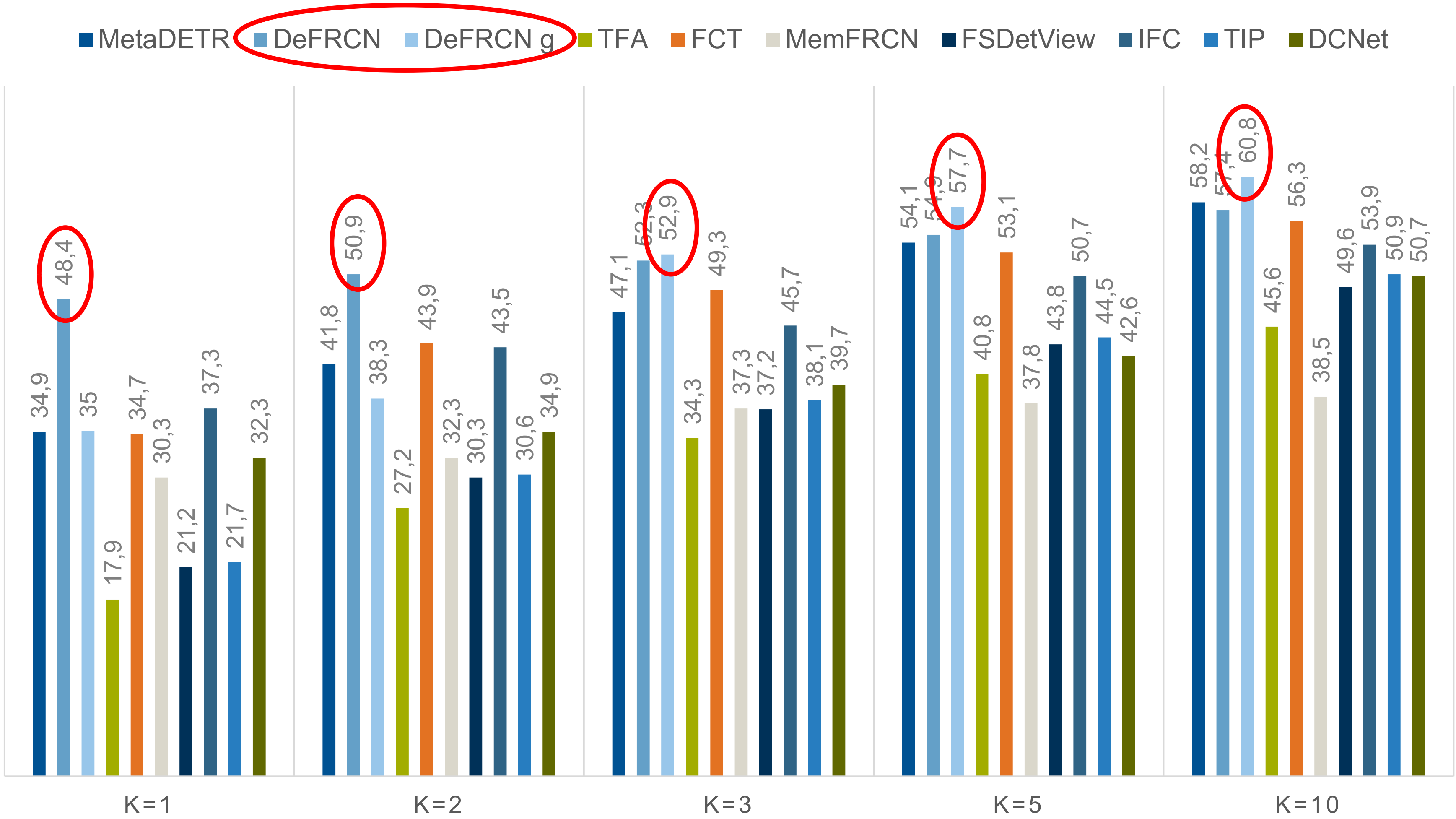
Results - Pascal Voc, Set 1 (Metric: AP_{50})



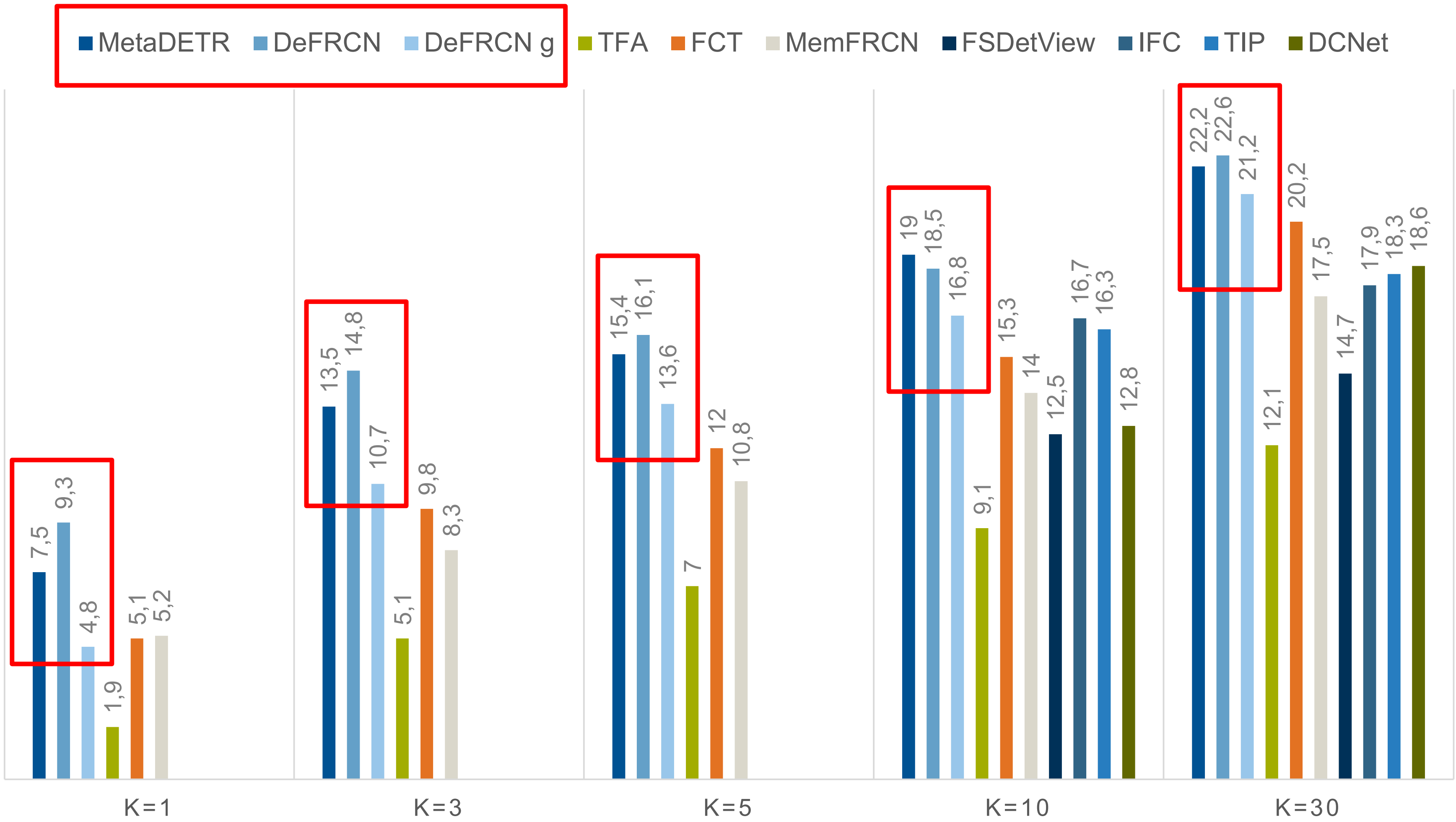
Results - Pascal Voc, Set 2 (Metric: AP_{50})



Results - Pascal Voc, Set 3 (Metric: AP_{50})



Results - Microsoft COCO (Metric $AP_{50:95}$)



References

[1] - Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, Shijian Lu, and Eric P. Xing. Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–12, 2022.

[2] - Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection, 2021.

[3] - Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-toend object detection with transformers, 2020.

[4] - Mona Kohler, Markus Eisenbach, and Horst-Michael Gross. ¨ Few-shot object detection: A comprehensive survey, 2021.

[5] - Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

Thank you for your attention

Alaa Arbi

Munich, 17. and 18. January 2023