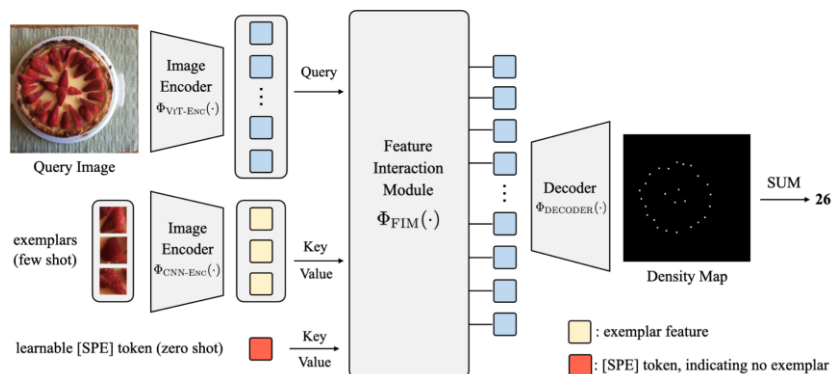


Generalised Visual Object Counting

The goal is to count the salient objects of **arbitrary** semantic class in an image, *i.e.* open-world visual object counting, with **arbitrary** number of “exemplars” provided by the end users, *i.e.* from zero-shot to few-shot object counting

Architecture of Counting Transformer(CounTR)



- **Visual Encoder**
 - ViT-based Query Image Encoder
 - CNN-based Exemplar Encoder
- **Feature Interaction Module**
 - Transformer Decoder Blocks
- **Visual Decoder**
 - Progressive Up-sampling Layers

Training Strategy

Two-stage Training Scheme

- Supervised Fine-tuning
- Self-supervised Pre-training with MAE

Scalable Mosaicing

Mosaicing: a scalable pipeline for synthesizing training images .



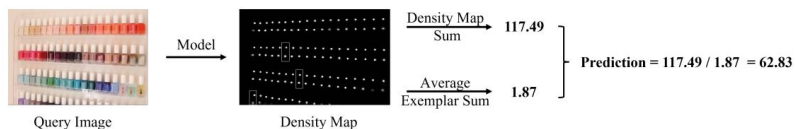
(a) Type A: using four images.

(b) Type B: using one image.

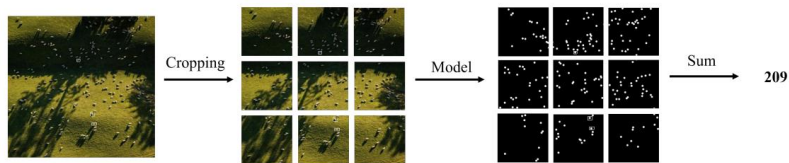
(1) stands for crop and scale, and (2) stands for collage and blending.

Test-time Normalisation

Test-time Normalisation: A strategy to calibrate the density map.



(a) Test-time Normalisation.



(b) Test-time Cropping.

Experiments

- **FSC-147** : A multi-class few-shot object counting dataset

| Methods | Year | Backbone | # Shots | Val | | Test | |
|----------------------|-----------------|-----------------|----------|--------------|--------------|--------------|---------------|
| | | | | MAE | RMSE | MAE | RMSE |
| RepRPN-C [11] | Arxiv2022 | ConvNets | 0 | 31.69 | 100.31 | 28.32 | 128.76 |
| RCC [5] | Arxiv2022 | Pre-trained ViT | 0 | 20.39 | 64.62 | 21.64 | 103.47 |
| CounTR (ours) | BMVC2022 | ViT | 0 | 17.40 | 70.33 | 14.12 | 108.01 |
| FR [7] | ICCV2019 | ConvNets | 3 | 45.45 | 112.53 | 41.64 | 141.04 |
| FSOD [1] | CVPR2020 | ConvNets | 3 | 36.36 | 115.00 | 32.53 | 140.65 |
| P-GMN [9] | ACCV2018 | ConvNets | 3 | 60.56 | 137.78 | 62.69 | 159.67 |
| GMN [9] | ACCV2018 | ConvNets | 3 | 29.66 | 89.81 | 26.52 | 124.57 |
| MAML [2] | ICML2017 | ConvNets | 3 | 25.54 | 79.44 | 24.90 | 112.68 |
| FamNet [12] | CVPR2021 | ConvNets | 3 | 23.75 | 69.07 | 22.08 | 99.54 |
| BMNet+ [15] | CVPR2022 | ConvNets | 3 | 15.74 | 58.53 | 14.62 | 91.83 |
| CounTR (ours) | BMVC2022 | ViT | 3 | 13.13 | 49.83 | 11.95 | 91.23 |

- **CARPK**: A class-specific car counting benchmark

| Methods | Year | CARPK | |
|----------------------|-----------------|-------------|-------------|
| | | MAE ↓ | RMSE ↓ |
| YOLO | CVPR2016 | 48.89 | 57.55 |
| Faster-RCNN | NIPS2015 | 47.45 | 57.39 |
| RetinaNet | ICCV2017 | 16.62 | 22.30 |
| IEP Count | TIP2018 | 51.83 | - |
| PDEM | CVPR2019 | 6.77 | 8.52 |
| GMN | CVPR2021 | 7.48 | 9.90 |
| FamNet | CVPR2021 | 18.19 | 33.66 |
| BMNet+ | CVPR2022 | 5.76 | 7.83 |
| CounTR (ours) | BMVC2022 | 5.75 | 7.45 |

- **Val-COCO & Test-COCO**: FSC-147 subsets from COCO

| Methods | Val-COCO | | Test-COCO | |
|----------------------|--------------|--------------|--------------|--------------|
| | MAE ↓ | RMSE ↓ | MAE ↓ | RMSE ↓ |
| Faster-RCNN | 52.79 | 172.46 | 36.20 | 79.59 |
| RetinaNet | 63.57 | 174.36 | 52.67 | 85.86 |
| Mask-RCNN | 52.51 | 172.21 | 35.56 | 80.00 |
| FamNet | 39.82 | 108.13 | 22.76 | 45.92 |
| CounTR (ours) | 24.66 | 83.84 | 10.89 | 31.11 |

- **Qualitative Results**

