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.

Experiments

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

| Methods       | Year | Backbone | # Shots | MAE   | RMSE   | MAE   | RMSE   |
|---------------|------|----------|---------|-------|--------|-------|--------|
| RepRPN-C [11]| 2022 | CoreNet  | 0       | 34.69 | 101.31 | 28.32 | 126.76 |
| RUC [5]       | 2022 | Pre-trained ViT | 0      | 30.39 | 60.82  | 21.64 | 105.47 |
| CounTR (ours) | 2022 | ViT      | 0       | 17.80 | 70.33  | 14.12 | 108.01 |
| FR [7]        | 2019 | CoreNet  | 3       | 45.45 | 112.53 | 41.64 | 141.04 |
| FSOD [10]     | 2020 | CoreNet  | 3       | 36.36 | 115.00 | 32.53 | 140.65 |
| P-GSM [9]     | 2018 | CoreNet  | 3       | 69.56 | 172.78 | 62.89 | 159.87 |
| GMN [9]       | 2018 | CoreNet  | 3       | 29.66 | 89.81  | 26.52 | 121.57 |
| MAM [12]      | 2017 | ICML2017 | 3       | 25.54 | 79.41  | 24.90 | 112.68 |
| FAMN [12]     | 2021 | CoreNet  | 3       | 23.75 | 69.07  | 22.98 | 99.34  |
| BMNet [15]    | 2022 | CoreNet  | 3       | 15.74 | 58.53  | 14.62 | 91.83  |
| CounTR (ours) | 2022 | ViT      | 3       | 13.13 | 49.83  | 11.58 | 91.23  |

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

| Methods       | Year | MAE | RMSE |
|---------------|------|-----|------|
| YOLO          | 2016 | 48.89 | 57.55 |
| Faster-RCNN    | 2015 | 47.45 | 57.39 |
| RetinaNet      | 2017 | 16.62 | 22.30 |
| FPEM          | 2018 | 51.83 | 51.83 |
| CVPR2019      | 2019 | 6.77  | 8.52  |
| GMN           | 2021 | 7.18  | 9.90  |
| FastNet       | 2021 | 18.19 | 33.66 |
| BMNet         | 2022 | 5.76  | 7.81  |

| Methods       | Year | MAE | RMSE |
|---------------|------|-----|------|
| FSC-147       | 2022 | 5.55 | 7.45  |

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

| Methods       | Year | MAE   | RMSE   |
|---------------|------|-------|--------|
| Faster-RCNN   | 2016 | 52.79 | 172.46 |
| RetinaNet     | 2017 | 63.57 | 174.36 |
| Mask-RCNN     | 2018 | 52.51 | 172.21 |
| FastNet       | 2019 | 39.82 | 108.13 |

| Methods       | Year | MAE   | RMSE   |
|---------------|------|-------|--------|
| FSC-147       | 2022 | 83.84 | 10.89  |

- **Qualitative Results**