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

**Results**

**CounTR**: Transformer-based Generalised Visual Counting

Chang Liu, Yujie Zhong, Andrew Zisserman, Weidi Xie

Project page: https://verg-avesta.github.io/CounTR_Webpage/

- **CARPK**: A class-specific car counting benchmark
- **Val-COCO & Test-COCO**: FSC-147 subsets from COCO

**Tasks**

- **FSC-147**
- **CARPK**
- **Val-COCO & Test-COCO**

**MAE**

- **Val-COCO**
- **Test-COCO**