Unsupervised Crowd Counting with CLIP

Morten Møller Christensen (s204258), Jacob Schrøder Ipsen (s204440). Frederik Emil Nagel (s204213)
DTU Compute, Technical University of Denmark

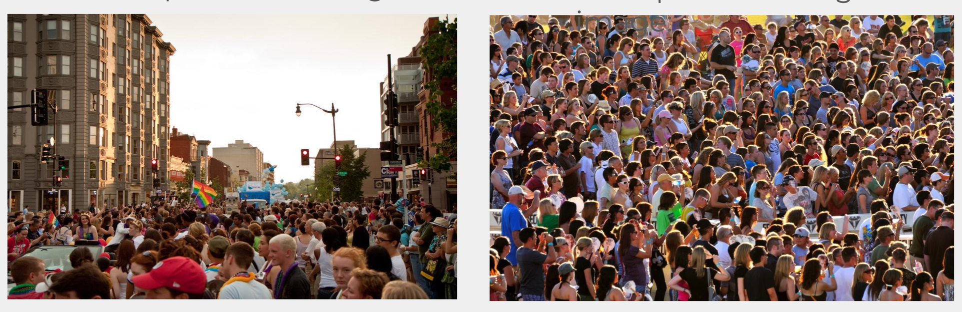
Introduction

- This project explores unsupervised crowd counting in images using a pre-trained CLIP model.

Dataset

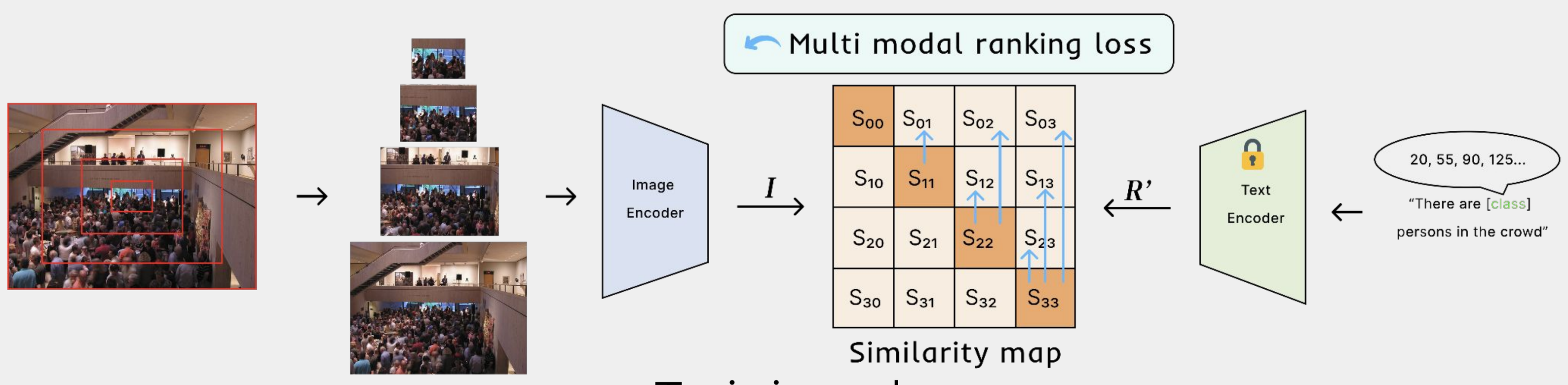
- 300 train images.
- 182 test images.
- Half of the test images are used for validation
- No transformations

Example of test image: Example of training

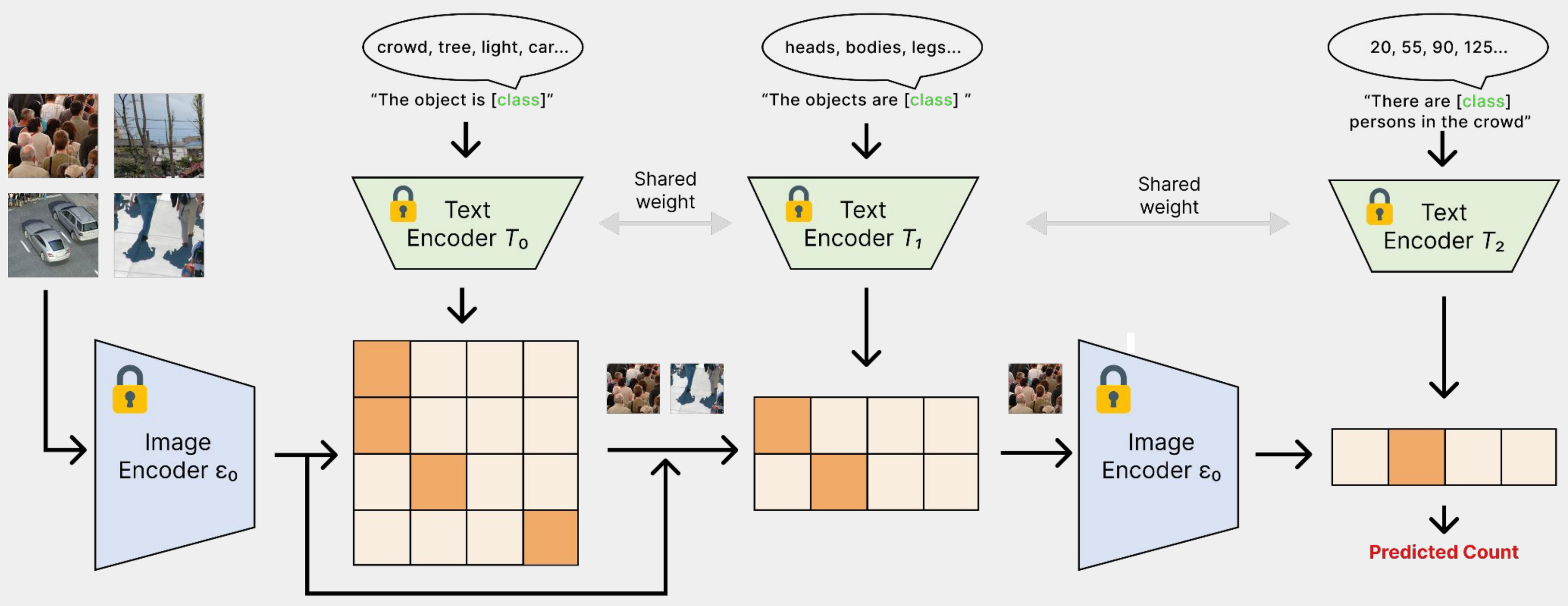


Baseline

- Progressively extracts larger image crops.
- Uses CLIP to generate feature vectors for both image crops and text prompts.
- Trains the image encoder with a multimodal ranking loss to align smaller crops with corresponding prompts.
- At inference, applies a two-step filtering process to discard non-crowded regions.



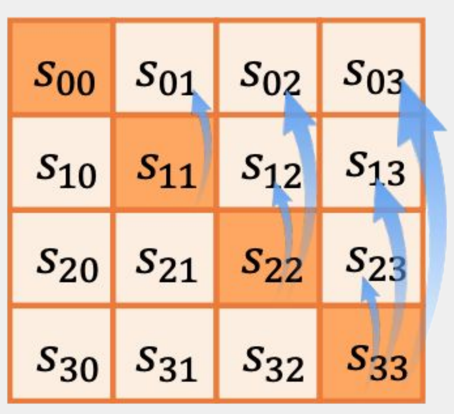
Training phase



Inference phase

Multimodal ranking loss (MMR)

$$L_r = \max(0, s_{j,i} - s_{i,i}),\ j < i$$



- $s_{a,b}$: Similarity between image patch **a** and text prompt **b**
- Enforces larger image patches align better with higher ranked prompts
- For <u>patch</u> **i** similarity to <u>prompt</u> **i** should be higher than to other prompts **j<i**
- Reflects assumption: larger patches contain more people

Increased Prompt Size: (IPS)

- Increased number of crop images from 6 to 10.
- Made more "diverse" prompt numbers.
- The core idea was to enable accurate prediction of smaller crowd counts.

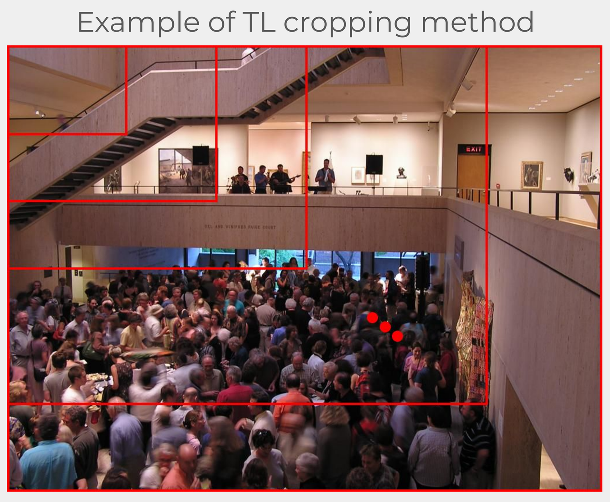
Original prompt numbers:
[20, 55, 90, 125, 160, 195]

New prompt numbers:
[5, 30, 55, 80, 105, 130, 155, 180, 205]

Top Left: (TL)

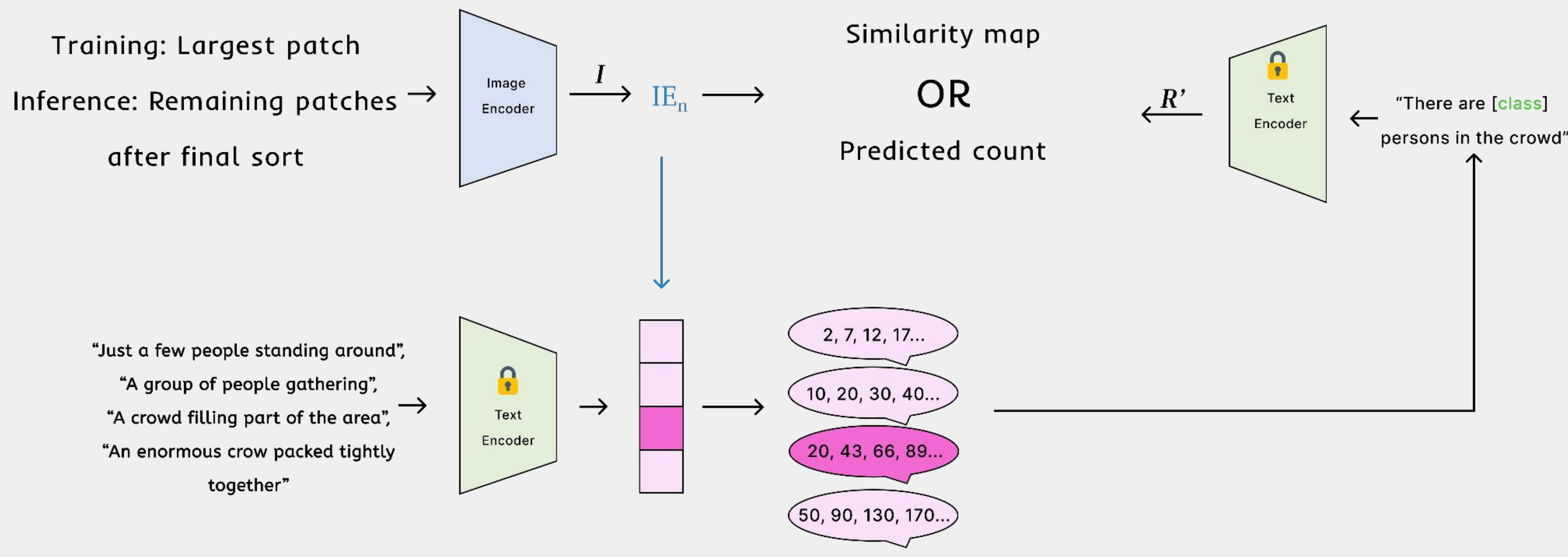
- Core idea was to explore a different way of cropping
- We crop from top left corner in increasingly bigger crops

Example of TL cropping method



Crowd Size Screening (CSS)

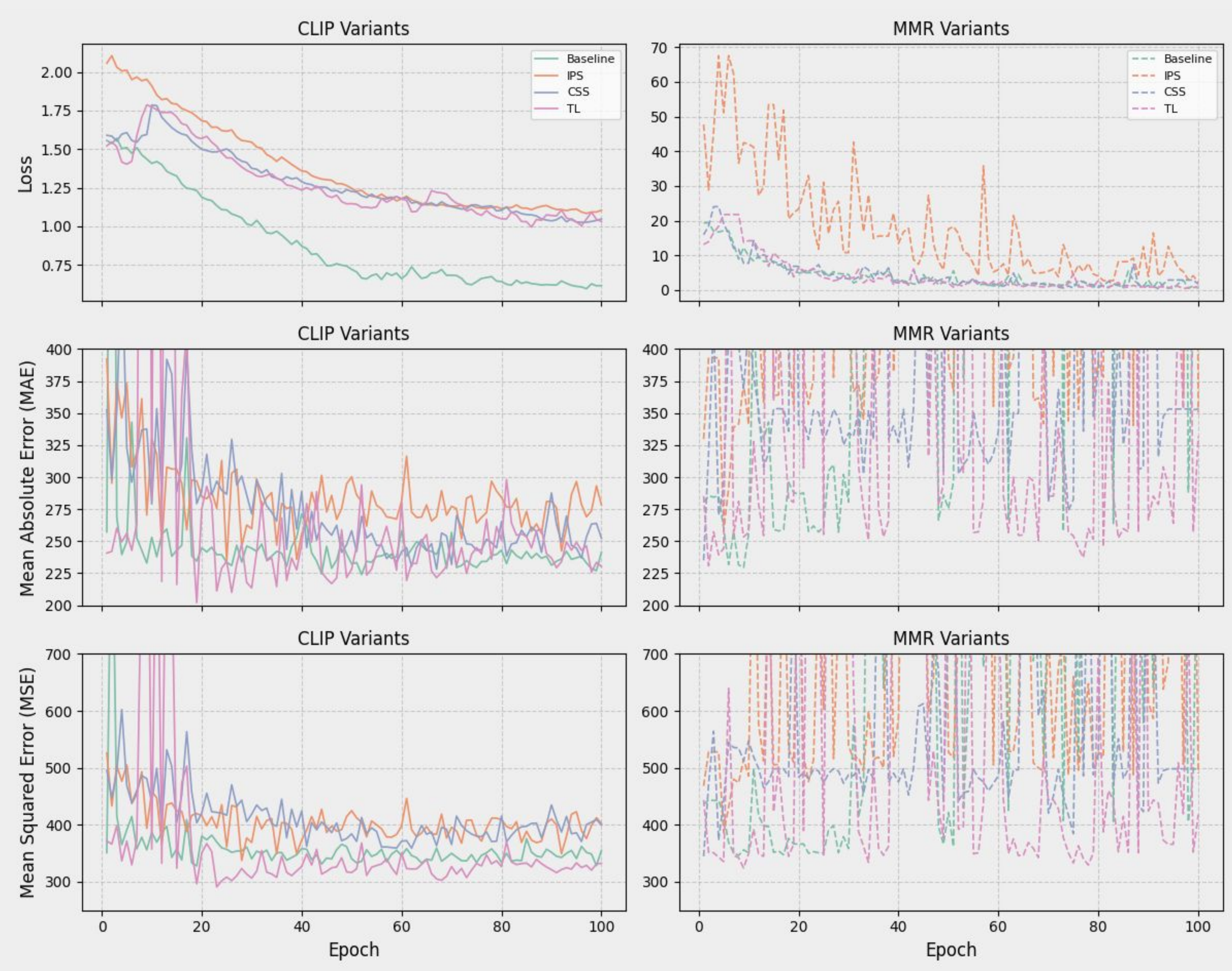
- Counts dynamically adjust based on crowd density of largest patch.
- Same design applied during inference (post-final sift step) to every remaining patch.
- Introduced number variety to improve performance.



Training:

- Batch size: 1
  lr: 1e-4
- Optimizer = RAdam

- MMR variants are highly unstable
- Large variance in MAE/MSE despite decreasing loss
- CLIP variants show smoother convergence
- Baseline (MMR) loss and MAE are poorly correlated



Results

- Lower training loss for MMR did not translate into lower MAE
- CLIP based models are more stable and accurate
- IPS and CSS outperform Baseline in both MAE and MSE
- All models underestimate large crowds significantly

Model	MSE	MAE
Baseline (MMR)	1029	937
Baseline (Clip)	355	241
IPS (MMR)	484	347
IPS (Clip)	402	277
CSS (MMR)	355	273
CSS (Clip)	349	232
TL (MMR)	418	331
TL (Clip)	**332**	**229**

Discussion and future work

- Prompt structure matters: Fixed sorting introduces bias
- Prompt tuning or distractor classes may reduce false positives
- Directly fine tuning the CLIP encoder is too aggressive: Use a lightweight adapter instead
- Explore alternative sorting strategies for inference ranking
  - Why not sort in other ways?