Improving Local Features with Relevant Spatial Information by Vision Transformer for Crowd Counting

Nguyen H. Tran, Ta Duc Huy, Soan T. M. Duong, Phan Nguyen, Dao Huu Hung, Chanh D. Tr. Nguyen, Trung Bui, Steven Q. H. Truong

Introduction

- Vision Transformer (ViT) and its variants demonstrate outstanding performance in many computer vision tasks.
- Current works employ ViT in overall image for crowd counting, which might not consistently focus on the crowd regions and is sensitive to errors under the circumstances of varied crowd densities and human scales.
- Considering above limitations to propose LoViTCrowd, our contributions are summarized as follows:
  i. We present a patch-based approach for crowd counting. Each sample is a patch comprised of 9 32 x 32 pixels cell, annotated by the respective human count of the central cell.
  ii. We marry CNN and ViT to construct the proposed LoViTCrowd that estimates the people in the central cell from the global context of the patch within which it resides.
- The proposed LoViTCrowd achieves state-of-the-art performance on four publicly available crowd counting benchmarks while being very simple to implement.

Datasets

Four public crowd counting datasets were used to evaluate the proposed method.

Datasets	ShangHai (A/ B)	UCF-QNRF	Mall
No. samples	1198	1535	2000
Ranges	[9, 578]	[50, 12000]	[5, 60]

Table 1: Distribution of the datasets.

Method

- We propose LoViTCrowd, a Transformer-based model with a patch embedding extractor followed by a regression head. With a 3 × 3 grid of 32 × 32 cells, target is to count the people in the central cell.

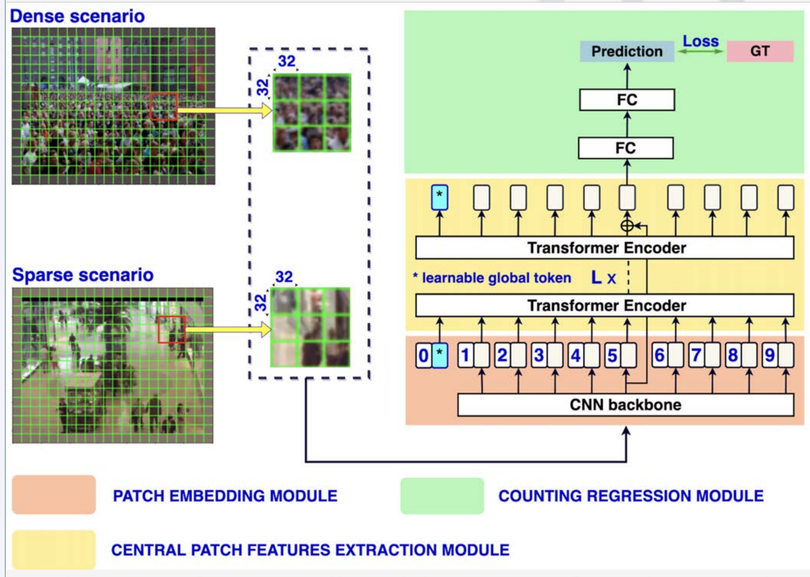

Figure 1: The overall architecture of LoViTCrowd.

- The estimated crowd number of an image is presented by summing all its local non-overlapping patches' counts.

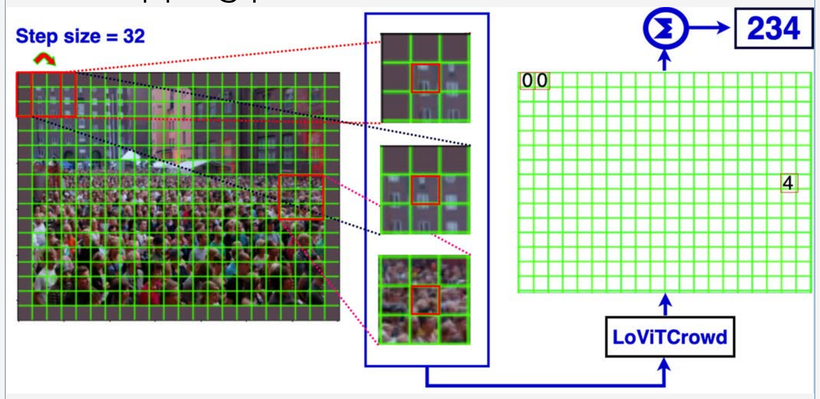

Figure 2: Crowd counting inference procedure.

Results

- LoViTCrowd vs. other SOTA methods:

Method	SHTech A		SHTech B		UCF-QNRF	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Sorting [38]	104.6	145.2	12.3	21.2	-	-
MATT [18]	80.1	129.4	11.7	17.5	-	-
TransCrowd-T [21]	69.0	116.5	10.6	19.7	98.9	176.1
TransCrowd-G [21]	66.1	105.1	9.3	16.1	97.2	168.5
CCTrans [32]	64.4	95.4	**7.0**	**11.5**	92.1	158.9
LoViTCrowd	**54.8**	**80.9**	8.6	13.8	**87.0**	**141.9**

Table 2: Performance of methods for crowd counting on SHTech Part A/B and UCF-QNRF.

Method	Mall	
	MAE	RMSE
Method in [14]	2.74	3.46
ConvLSTM-nt [37]	2.53	11.2
ConvLSTM [37]	2.24	8.5
Bi-ConvLSTM [37]	2.10	7.6
TransCrowd-G [21]	1.72	2.18
LoViTCrowd	**1.66**	**2.10**

Table 3: Performance of methods for crowd counting on Mall.

Figure 3: Visualization on a UCF-QNRF's test sample.

Code available at: https://github.com/nguyen1312/LoViTCrowd