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

BMVC 2022 Submission # 729

This supplementary material provides a more detailed description of LoViTCrowd, including data preparation and network architecture. Additional experiments are provided to analyze the impacts of the central cell's neighborhood in crowd counting performance. Beside the quantitative result in the manuscript, we present some qualitative results here for clarity purpose.

1 Implementation details

1.1 Data preprocessing

We conducted experiments on publicity datasets, i.e., Mall, ShangHaiTech Part A, ShangHaiTech Part B, UCF-QNRF. Each data sample is annotated with the 2d-coordinate of the human head. The crowd estimation groundtruth of each 96×96 patch is the number of head in the central 32×32 cell. Those image patches are extracted by using sliding window as described in the main paper.

1.2 Network architecture

In patch embedding module, instead of using raw image patches, we followed the hybrid architecture described in [2] and employed the ImageNet [1] pre-trained ResNet34 [3]. We used the output feature maps generated from the fourth stage as inputs for the Transformer encoder. Due to the difference between the input image resolution for our proposed method, i.e., 96×96, and the original one in the pre-trained ViT [2] we used, i.e., 384×384, we adapted the bi-linear 2D interpolation introduced in [4] in order that we could load to fit the dimensions of the pretrained weight for position embedding layer.

2 Additional experiments

To further investigate the importance of the adjacent cells, we experimented with more configurations and measured the crowd counting performance. For instance, as visualized in Fig. 1 and Fig. 2, we mask the subset of the eight neighboring cells before making prediction in the central one. Table 1 summarized the results of six experiments. As shown in table 1, To estimate the number of people in central cell, LoVitCrowd relies on the context of the neighboring cells which are cross the central one, especially the bottom neighboring cells. For

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

Figure 1: Masking four cells in the four corners (left) and masking four cells in cross positions (right) before predicting the number of people in central cell (red bounding box areas).

Figure 2: Masking a set of three consecutive cells in four ways, i.e., top, bottom, left, right, respectively before predicting the number of people in central cell (red bounding box areas)

instance, when masking four cells across the central one, the counting errors significantly increase, i.e., 27.7 in MAE, 44.6 in RMSE. When masking three below neighboring cells, the MAE is up to 33.3, and the RMSE is 47.3.

Table 1: Performance of our proposed LoViTCrowd on ShangHai Tech Part B on six neighboring cells' masking experiments.

Cells'masking configuration	MAE	RMSE
Corner	10.2	16.2
Cross	27.7	44.6
Top	18.3	22.6
Left	10.9	16.3
Right	9.5	14.8
Bottom	33.3	47.3

3 Qualitative result

To show the qualitative evaluation on four public datasets, we visualized four respective examples including ground truths, prediction values and the attention maps on several 96 $\times$ 96 patches from our proposed LoViTCrowd, i.e., fig. 3, 4, 5 and 6. From the central cell's query, our approach generates reasonable attention weights on eight adjacent cells.

Figure 3: Qualitative visualization on a Mall's test sample.

With the help of relevant contextual spatial information from neighboring cells captured by ViT, LoViTCrowd obtains robust performance across various datasets with different crowd scenarios.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

Figure 4: Qualitative visualization on a ShangHai Tech Part A's test sample.

Figure 5: Qualitative visualization on a ShangHai Tech Part B's test sample.

Figure 6: Qualitative visualization on a UCF-QNRF's test sample.