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

Nguyen H. Tran[1]
v.nguyentran@vinbrain.net

Ta Duc Huy[1]
v.huyta@vinbrain.net

Soan T. M. Duong[1,3]
v.soanduong@vinbrain.net

Phan Nguyen[1]
v.nguyenphan@vinbrain.net

Dao Huu Hung[1]
v.hungdao@vinbrain.net

Chanh D. Tr. Nguyen[1,2]
v.chanhng@vinbrain.net

Trung Bui
bhtrung@yahoo.com

Steven Q.H. Truong[1]
v.brain01@vinbrain.net

[1] VinBrain JSC.,
Vietnam

[2] VinUniversity,
Vietnam

[3] Le Quy Don Technical University,
Vietnam

Abstract

Vision Transformer (ViT) variants have demonstrated state-of-the-art performances in plenty of computer vision benchmarks, including crowd counting. Although Transformer based models have shown breakthroughs in crowd counting, existing methods have some limitations. Global embeddings extracted from ViTs do not encapsulate fine-grained local features and, thus, are prone to errors in crowded scenes with diverse human scales and densities. In this paper, we propose LoViTCrowd with the argument that: LOcal features with spatial information from relevant regions via the attention mechanism of ViT can effectively reduce the crowd counting error. To this end, we divide each image into a cell grid. Considering patches of 3 × 3 cells, in which the main parts of the human body are encapsulated, the surrounding cells provide meaningful cues for crowd estimation. ViT is adapted on each patch to employ the attention mechanism across the 3 × 3 cells to count the number of people in the central cell. The number of people in the image is obtained by summing up the counts of its non-overlapping cells. Extensive experiments on four public datasets of sparse and dense scenes, i.e., Mall, ShanghaiTech Part A, ShanghaiTech Part B, and UCF-QNRF, demonstrate our method's state-of-the-art performance. Compared to TransCrowd, LoViTCrowd reduces the root mean square errors (RMSE) and the mean absolute errors (MAE) by an average of 14.2% and 9.7%, respectively. The source is available at https://github.com/nguyen1312/LoViTCrowd.

1 Introduction

Transformers [34] are famous for their state-of-the-art performance on many natural language processing tasks. For the last two years, it has been adapted to computer vision domains and achieved superiority over convolutional neural networks (CNNs) when trained on large-scale datasets. Vision Transformer (ViT) and its variants, without using any convolutional layer, demonstrate outstanding performance in image classification [10]. [10][24][8] have shown that Transformer-based models can learn the discriminative features between distinct image patches more effectively than CNNs. To the best of our knowledge, TransCrowd [21] is the first work to employ ViT for crowd counting. The image is permuted into flattened 2D patches, which are fed to Transformer encoders for the global context feature. It comes with a regression head to estimate the total number of people using this global context feature. TransCrowd employs the attention mechanism and learns the global context feature as the patches interact. We argue that the correlation across the image patches does not contribute to the crowd counting performance. Each person only takes up a small area in the patch; thus, other patches bring no relevant spatial information to the current patch. Besides, TransCrowd captures the global context from too large regions of the images (Fig. 1),

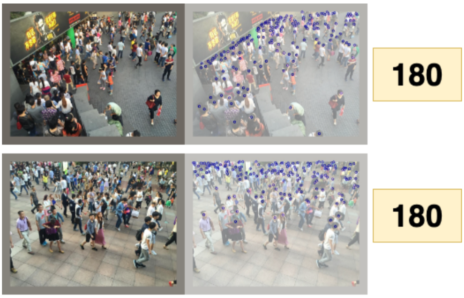

Figure 1: Despite the same number of people, the human scales and crowd densities are different between two cases.

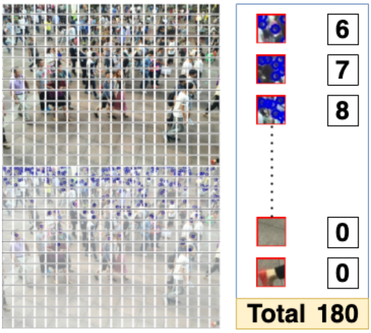

Figure 2: LoViTCrowd estimates the number of people in each patch of the whole image.

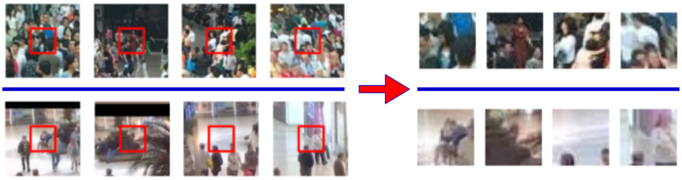

Figure 3: In dense view, because people are captured by large-scale cameras, most human heads are small. Considering the central 32 × 32 cells (red bounding boxes), they are usually fully encapsulated. In sparse scenarios, people are often captured by close-range cameras. The majority of human heads in the regions of interest span several cells.

which might contain unnecessary information. Therefore, it does not consistently focus on the crowd regions and is sensitive to errors under the circumstances of varied crowd densities and human scales.

In this paper, we consider the above limitations to propose LoVitCrowd. We divide an image into small patches of fixed size (e.g., 32 × 32 pixels) as shown in Fig. 2, ensuring each patch still captures the visual properties corresponding to the scene condition of the whole image. Each cell can contain one or more heads depending on the camera perspective (Fig. 3). The surrounding cells in a patch, which capture other main parts of a human body, provide contextual information to estimate the number of people in the central cell. Cells from other patches are irrelevant to the central cell, thus, are left unconsidered. Instead of using image-level tokens, we utilize patch-level representation, which was efficiently adopted for dense prediction tasks, i.e., object detection [2], object segmentation [36], and crowd counting. Our contributions are summarized as follows:

(1) We present a patch-based approach for crowd counting. Each sample is a patch comprised of 9 32 × 32 pixels cell, annotated by the respective human count of the central cell.

(2) Instead of flattening raw image patches, 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 results on four public datasets, i.e., Mall, ShangHai Tech Part A, ShangHai Tech Part B, and UCF-QNRF. We also conducted an extensive set of ablation experiments to provide insights into several configurations of LoViTCrowd, including cross-domain evaluation, training volume and resolution variations, and the permutation importance of the adjacent cells.

2 Related works

In the early literature, detection-based approaches [19][9][11][33], even related approaches using recent state-of-the-art detectors, i.e., Faster R-CNN [28], YOLO [27], RetinaNet [22], etc., seemed to perform poorly in high-density crowds with heavy occlusion. Regression-based methods were introduced to improve the counting performance. Several methods [4][3][16] extracted the low-level features of the scene for counting regression. They are likely to produce unsatisfactory results.

CNN based methods can handle scene adaptation and scale diversity issues. Related approaches interpret the count number directly or via density map estimation. To address the issue of multi-scale scenes, MCNN [40], Switch-CNN [1], CAN [23] incorporated multi-size filters to extract multi-scale features. [29] introduced a single-branch CNN that learns two tasks simultaneously, i.e., counting classification and density map regression. [37] incorporated CNN with long-short-term memory (LSTM) [15] to capture spatial-temporal information for crowd counting. CSRNet [20] adapted dilated convolutional layers to improve the output quality. A convolutional LSTM model [14] was introduced to interpret the people to count in every image's local 32 × 32 patches with the help of their eight respective neighboring cells. The sorting network [38] was proposed for directly regressing counting without location-level annotations. P2PNet [30] directly received point-level annotations as its learning targets. The model MATT [18] was designed for crowd counting with a small number of location-level annotations and a large number of count-level annotations. Conventional CNNs often use a down-sampling mechanism to generate large receptive fields in their deeper layers, which causes a reduction in spatial resolution. Moreover, CNN based approaches use convolutional kernel's receptive field with limited size. Therefore, they fail to capture global context information and discard local semantic information, which is crucial to crowd counting. If the approach's concept is only regression, the counting performance is often limited due to the quality of image features extracted by CNNs. Some approaches [37] [14] try to boost the crowd counting performance by adapting the combination of CNN and LSTM to explore the local spatial context incorporation.

Transformer [34] has been shown to be better than LSTM in terms of performance and computational efficiency. Transformer has demonstrated revolutionary performance improvement in various computer vision tasks, such as image classification [10], object detection [2], and object segmentation [36]. Inspired by Vision Transformer, recent works use the Transformer for crowd counting where TransCrowd [21] is the pioneer. Given an image, after extracting the image's global embedding, it directly regresses the number of people in two ways, i.e., with the proposed TransCrowd-Token and TransCrowd-GAP. Another work, CCTrans [32], adopts a pyramid transformer to extract multi-level feature maps for learning targets. Both TransCrowd and CCTrans do not exploit data enrichment, while Transformer-based methods need a large-scale dataset. Furthermore, existing approaches output the number of people from spacious areas via global-context visual features, which is not always effective because of the diversely specified crowd distribution in the scene.

Following the previous work [14] as mentioned above and using the power of the Transformer in computer vision, we formulated the task as crowd counting in every cell of the image grid. Such framework is simple to implement but achieve comparable results, compared to methods that estimate the people counting from the whole image. We developed a Transformer-based model with a patch embedding extractor followed by a regression head for crowd counting, named LoViTCrowd. LoViTCrowd utilizes the Transformer-encoder from [10] to every 32 × 32 cells with its surrounding cells' collaboration in an image grid to capture robustness embedding for counting. This strategy enriches training data samples spectacularly because, from an image, we can slice it into as many smaller patches, including overlapping and non-overlapping ones.

3 Methodology

As depicted in Fig. 4, LoViTCrowd includes three modules: (1) the patch embedding module, (2) the central patch's features extraction module, and (3) the counting regression module. With a 3 × 3 grid of 32 × 32 cells, our target is to count the people in the central cell. Each 3 × 3 grid is encoded with the patch embedding module into as a sequence of 9 D-dimensional embeddings. The central patch feature extraction module consists of a stack of Transformer-encoders. The module extracts the features of the fifth cell (the central cell) with self-attention mechanism, leveraging spatial contextual information from its eight surrounding cells. In addition, for crowded scenarios, restricting the receptive field for each considered query 32 × 32 cell to its neighboring cells avoids learning redundant and irrelevant information for the task of crowd counting. We adopt skip-connection [12] from the central patch embedding to its final layer feature to consolidate the visual context information of its respective region. Finally, the counting regression module consists of two fully connected layers to estimate the number of people within the central cell.

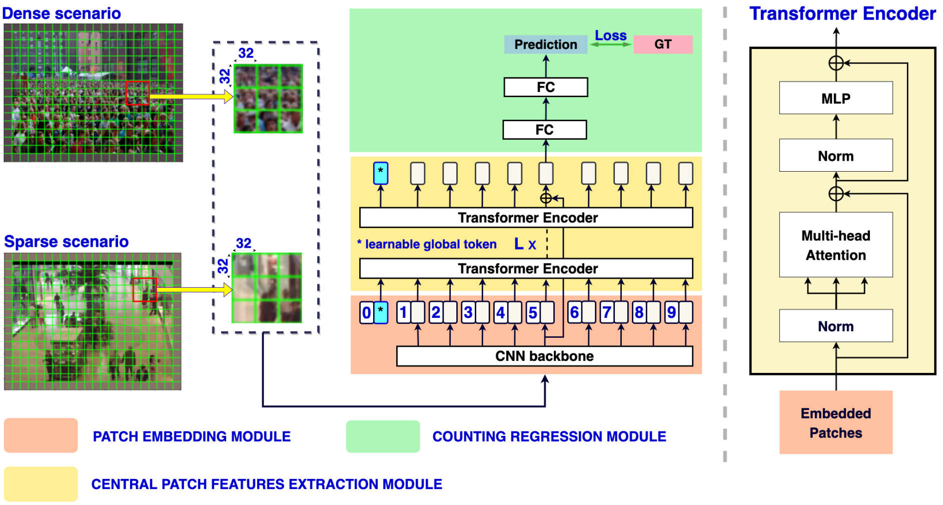

Figure 4: The overall architecture of the proposed LoViTCrowd model.

3.1 Data pipeline

Since the 32 × 32 cells in the boundary area do not have enough eight respective adjacent neighbors, every frame's boundaries are padded with the image's mean value for 32 pixels for each size (Fig. 4). We divide each whole frame into many sub-images of size 96×96 via a sliding window.

Given a step size s, each input image with a resolution $H \times W$ is split into N patches as follow:

$$N = \left\lfloor \frac{H+s-96}{s} \right\rfloor \times \left\lfloor \frac{W+s-96}{s} \right\rfloor \tag{1}$$

Smaller stride s leads to larger number of patches N. In this study, for training phase, we choose $s = 32$ and $s = 10$. The number of training samples is $N_{total} = N_{s=10} + N_{s=32}$.

3.2 Patch embedding module

CNN feature extraction. As shown in Fig. 5, given an image $x \in R^{96\times96\times3}$, the patch embedding module employs the Imagenet pretrained Resnet34 [12] to extract visual crowd features. We obtain the output of the last Resnet34's block, i.e, $F \in R^{H\times W\times C}$ (H, W, C are the height, width, and channel size of the feature maps) and divide the feature map into nine feature patches with the size of $\frac{\sqrt{HW}}{3} \times \frac{\sqrt{HW}}{3} \times C$. Each feature patch is flattened into vectors $x \in R^{1\times D}$, where $D = \frac{HW}{9} \times C$ followed by a a linear projection $f : x_i \in R^{1\times D} \rightarrow e_i \in R^{1\times D'}$ ($D' = 768$ in our experiment settings).

Position embeddings. Like ViT, we incorporate learnable position embedding into each image token to retain positional information. This process could be formulated as: $I_{input} = [e_1 + p_1; e_2 + p_2; ...; e_9 + p_9]$, where $p_i \in R^{1\times D'}$, $i = 1, 2, ..., 9$.

3.3 Central patch's feature extraction module

Transformer-encoder. We adopt the standard Transformer encoder with the stack of $L = 12$ layers as the primary feature extractor. We suppose the input sequence of the l^{th}-th encoder layer is $Z_{l-1} = [Z_{l-1}^0, ..., Z_{l-1}^9]$, where the Z_l^0 is the special token to learn the global context at layer l. The main advantage of adopting the Transformer block for extraction is the multi-head attention mechanism that allows the tokens to interact and determine which tokens they should pay more attention to in the sequence. The multi-head attention mechanism is based on scale dot-product [34] scoring scheme, allowing the tokens to interact and pay attention to the most relevant region. The attention block is followed by a feed-forward layer with GeLU activation function [13] and dropout [31].

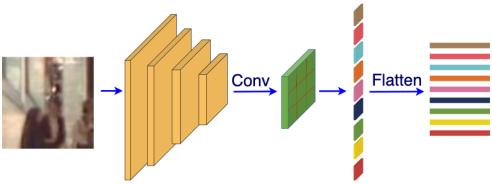

Figure 5: Patch embedding module employed the pretrained ResNet34 to capture deep visual patch features.

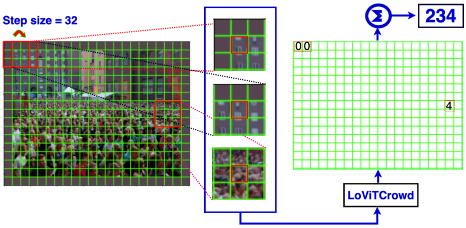

Figure 6: Crowd counting inference procedure.

Central patch's feature extraction. In the output sequence of the final encoder block $[Z_L^1, Z_L^2, ..., Z_L^9]$, we only consider the central patch representation Z_L^5 to estimate the crowd number. The fifth patch feature at the first layer Z_0^5 is added to Z_L^5 to obtain the central 32 × 32 cell embedding $z = Z_0^5 + Z_L^5$. z is then fed to the two non-linear fully connected layer of the counting regression module to estimate the number of people in the central cell.

4 Experiments

4.1 Implementation Details

We evaluate our approach across four benchmarks: ShangHaiTech (Part A/B) [39], UCF-QNRF [17], and Mall [5][7][25][6]. Except for Mall, whose images have the same size of 640 × 480, other datasets have various resolutions among their data samples. Therefore, we resize all the images in ShangHaiTech Part A/B and UCF-QNRF to the size of 1024 × 768.

Loss function. Euclidean distance (L2) is commonly used to train crowd counting models for its simplicity and robustness. Therefore, we choose L2 loss for calculating the error.

Training details. LoViTCrowd is implemented in PyTorch [26] and trained with a system having a NVIDIA A100-SMX4 GPU with 40 GB of memory. We used the pretrained ViT-B/32 provided by [10] as the backbone network for feature extraction. During training, we used a learning rate of 1^{e-4}. Adam optimizer was applied with default setting in beta1, beta2, epsilon (0.9, 0.999, 1^{e-8}, respectively) and 5^{e-4} in weight decay.

Counting people in the single image. As shown in Fig. 6, a padded image with a resolution of H × W is first divided into a grid of 32 × 32 cells. The estimated crowd number of an image is presented by summing all its local non-overlapping patches' counts. Therefore, a 96 × 96 sliding window moves from left to right and top-to-bottom with a stride

Table 1: Performance of methods for crowd counting on SHTech Part A/B and UCF-QNRF.

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 Mall.

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**

of 32 to cover every central cell and its eight surrounding cells. Followed by Eq. 1, we have total $N_{s=32}$ number of patches. Such 3 × 3 grids of cell are fed into LoViTCrowd to estimate the number of existing people in the central cell. The sum of people in $N_{s=32}$ patches is the final crowd number estimation.

4.2 Comparisons with State-of-the-art

Evaluation metrics. To evaluate the crowd counting performance, we used two standard regression metrics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

Results. Table 1 compares the proposed LoViTCrowd with existing crowd counting methods, including CNN based approach [38] [18], and Transformer based approach [21] [32], on ShangHaiTech, i.e., ShangHaiTech part A, ShangHaiTech part B, and UCF-QNRF. As shown in Table 1, LoViTCrowd achieves remarkable results in both MAE and RMSE. In ShangHaiTech Part A, our method shows a substantial improvement over other approaches, including the two most recent state-of-the-art approaches utilizing Transformer, i.e., TransCrowd and CCTrans. Compared to TransCrowd and CCTrans, LoVitCrowd decreases the MAE by up to 17.1% and the RMSE by up to 23.0%. Meanwhile, on ShangHaiTech Part B, compared to TransCrowd, LoViTCrowd is very competitive with an improvement of 7.5% in MAE and 14.3% in RMSE. However, the performance of LoViTCrowd is relatively poor compared to CCTrans. CCTrans adopted the Pyramid Vision Transformer (PVT) [35] for feature extraction, that is shown to be better than ViT for downstream tasks in the paper. Since the ShangHaiTech Part B benchmark has diversified scales, we find that pyramid architecture shows more favorable results than vanilla ViT. It will be the following research.

On the UCF-QNRF, one of the most challenging benchmarks for crowd counting tasks, we also achieve state-of-the-art performance compared to other approaches. Table 1 shows that our LoViTCrowd achieves 5.5% MAE and 10.7% RMSE improvement over CCTrans, the most recent state-of-the-art Transformer-based approach. We also do experiment with TransCrowd-GAP and the proposed LoViTCrowd on Mall, an extremely sparse crowd dataset. To make a fair comparison, we compare our method to the most recently implemented ones on Mall, i.e., [37] [14]. In Table 2, our method reduces the MAE by 39.4%, 21% 3.5%, and the RMSE by 39.3%, 72.4% and 3.7%, compared to [14], [37] and TransCrowd, respectively. In general, LoViTCrowd outperforms most previous crowd counting state-of-the-arts on two of the most common evaluation metrics, i.e., MAE and RMSE, across many datasets under various conditions.

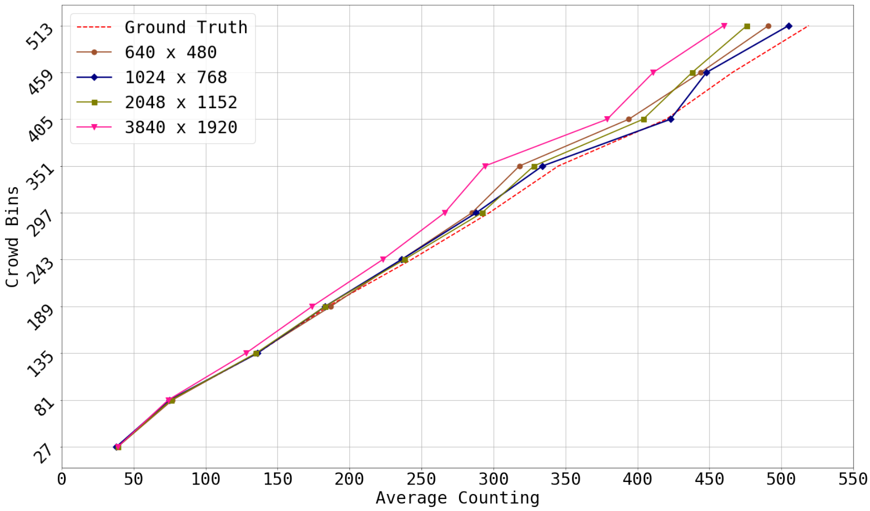

Figure 7: The comparison of the groundtruth and the estimated count results of different resolutions on ShangHai Tech Part B. The images are grouped into 10 bins.

4.3 Ablation Study

We conduct several ablation experiments to study the performance of different configurations, thus suggest some practical choices of LoViTCrowd for adoption.

Cross domain testing. In reality, the testing crowd scenarios are not always similar to the training ones. To evaluate the generalization of the LoViTCrowd, we conducted cross-domain evaluation where we train the model on one dataset and test on another dataset. Moreover, to learn the underlying crowd distribution rather than being overfitting to any specific dataset, we gathered all the training samples from four datasets to form a multi-domain dataset and train the LoViTCrowd using the same configuration.

Table 3: Performance of the proposed LoViTCrowd in cross-domain evaluation.

Pretrained	**SHTech A**		**SHTech B**		**UCF-QNRF**		**Mall**	
	MAE	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
SHTech Part A	**54.8**	**80.9**	24.2	64.3	212.4	418.6	3.3	4.1
SHTech Part B	126.7	216.8	**8.6**	**13.8**	270.4	491.7	3.1	3.9
UCF-QNRF	87.8	162.4	14.6	26.7	**87.0**	**141.9**	3.4	4.3
Mall	185.5	291.6	35.0	59.9	404.4	709.0	**1.7**	**2.1**
Multi-domain	68.0	116.2	11.3	18.4	93.4	152.9	2.5	3.1

As shown in the Table 3, on ShangHai Tech Part A/B and UCF-QNRF, the counting errors significantly increase when evaluated on a different crowd distribution. Our proposed model, The model trained on the multi-domain dataset achieves remarkable performance. It reaches the second place compared with different cross-domain configurations, this shows that adding many distributions to the training dataset will significantly improve the model generalizability.

Table 4: Performance of the proposed LoViTCrowd on Mall with the configuration of different sets of step sizes used in training phase.

Step Size	No. Train Samples	MAE	RMSE
s = {32}	240000	1.9	2.5
s = {10, 32}	2436000	**1.7**	**2.1**
s = {10, 20, 32}	3006400	**1.7**	2.2

Table 5: Performance of the proposed LoViTCrowd on ShangHai Tech Part B with the configuration of different resolutions.

Resolution	No. Train Samples	MAE	RMSE
640 × 480	966400	9.4	15.3
1024 × 768	2822800	8.6	13.8
2048 × 1152	9208800	**8.0**	**13.0**
3840 × 1920	30276800	10.7	17.8

Different choices of the step values for extracting patches in training. From Eq. 1, given a set of step value, i.e., $s = \{s_0, s_1, ..., s_K\}$, the volume of training dataset equals to $N_{total} = \sum_{i=0}^{K} N_{s=s_i}$. When using a 96 × 96 sliding window with varying step sizes over the image to extract patches for training, the result is further improved since the training dataset becomes larger. As shown in Table 4, with strides of $\{10, 32\}$, the counting errors on Mall decrease significantly, i.e., 10.5% MAE and 16.0% RMSE, compared to single-step size with a value of 32. The result of LoViTCrowd saturates when adapting more strides, i.e., 10, 20, and 32. Considering the trade-off between the total number of training samples and counting errors, the step value $s = \{10, 32\}$ is optimal.

Different resolutions. Higher resolution not does only lead to more patches, followed by Eq. 1, but it also "zooms in" the human scales for better crowd estimation result. Table 5 shows the ablation study of image's resolution on ShangHai Tech Part B, i.e., 640 × 480, 1024 × 768, 2048 × 1152 and 3840 × 1920. For a fair comparison, we conducted experiments with step values of s = $\{10, 32\}$ during training. The MAE and RMSE are reduced from 9.4 and 15.3 to 8.0 and 13.0, respectively, by resizing the images to 2048 × 1152. The counting errors are slightly improved compared to the resolution 1024 × 768. Too high resolution makes the cell grid not large enough to fully encapsulate the main visible human body, drastically reducing the quality of training samples. For instance, despite the enormous volume of data generated by initially resizing the images to 3840 × 1920, the counting performance decreases significantly. For more details, we visualize the comparison between actual counts and predicted count results from different resolution configurations in Fig. 7. ShangHai Tech Part B images are grouped into ten bins according to the groundtruth number of people in each sample. The y-axis is the average human counts of images in each bin. Considering the trade-off between computing cost and performance, the resolution of 1024 × 768 is recommended.

Permutation importance of the adjacent cells. When predicting the number of people in the central cell, we aggregate the context information from its neighboring cells. To highlight the importance of those cells, we mask one of the eight adjacent cells in each experiment and measure the performance on the whole ShangHai Tech Part B dataset as visualized in Fig. 8.

Fig. 9 shows the total MAE in each of the eight settings. The central cell shows the MAE of the original configuration, where no cell is masked. The higher the MAE indicates, the more important the cell as it is masked during testing.

Interestingly, when masked, the two top left and top right cells show an improvement in MAE. It suggests that those cells are not important to estimate the number of people in the central cell. Furthermore, masking those cells intuitively removes irrelevant information and improves crowd estimation performance.

Relevant information for the central cell are shown to reside in the second and the third rows as the performance decreases when masked. The cell below the centroid is remarkably important. Masking this cell leads to a surge in MAE. Considering a human body, this cell would contain the bodies of the humans, thus making it crucial to estimate the crowd number of the central cell where their heads are inside.

To explain the importance of the middle top cell that also shows a high MAE when masked, we anticipate that this cell would help the model avoid false positive head counts in the central cell. The central cell would not consider the bodies as people to be counted if their heads are within the top cell.

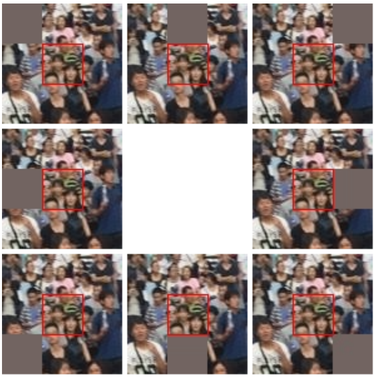

Figure 8: Eight different types of masking setting before predicting the people count in the central cell (red bounding box areas).

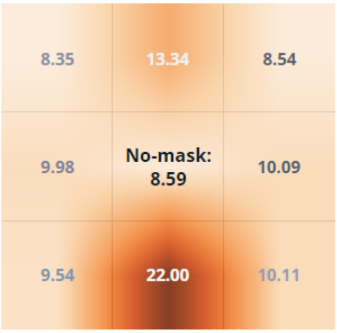

Figure 9: Visualization of LoViTCrowd's performance (MAE) on ShangHai Tech Part B with the respective configuration of neighboring cell's masking.

5 Conclusion

In this paper, we proposed LoViTCrowd, a novel cell-based network using ViT for crowd counting. Our model is designed to capture fine-grained features from every 32 × 32 cells so that it can effectively estimate the number of people locally. We conducted extensive experiments on four publicly available crowd counting benchmarks to demonstrate the superior performance of our proposed LoVitCrowd compared to several existing methods for crowd counting while being very simple to implement. Ablation studies are also carefully conducted to give insights into practical considerations of our method. We plan to evaluate our approach on other datasets with various crowd scenarios to justify its robustness in multiple domains.

References

[1] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5744–5752, 2017.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In *2009 IEEE 12th international conference on computer vision*, pages 545–551. IEEE, 2009.

[4] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–7. IEEE, 2008.

[5] Chen Change Loy, Shaogang Gong, and Tao Xiang. From semi-supervised to transfer counting of crowds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2256–2263, 2013.

[6] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *Bmvc*, volume 1, page 3, 2012.

[7] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2467–2474, 2013.

[8] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021.

[9] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011.

[10] 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.

[11] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[12] 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.

[13] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[14] Anh Hoang, Doan-Phuc Phan, Huu-Hung Dao, Van-Nam Huynh, et al. Human density estimation by exploiting deep spatial contextual information. In *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–5. IEEE, 2019.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013.

[17] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–546, 2018.

[18] Yinjie Lei, Yan Liu, Pingping Zhang, and Lingqiao Liu. Towards using count-level weak supervision for crowd counting. *Pattern Recognition*, 109:107616, 2021.

[19] Bastian Leibe, Edgar Seemann, and Bernt Schiele. Pedestrian detection in crowded scenes. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 878–885. IEEE, 2005.

[20] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[21] Dingkang Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. Transcrowd: weakly-supervised crowd counting with transformers. *Science China Information Sciences*, 65 (6):1–14, 2022.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[23] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5099–5108, 2019.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[25] Chen Change Loy, Ke Chen, Shaogang Gong, and Tao Xiang. Crowd counting and profiling: Methodology and evaluation. In *Modeling, simulation and visual analysis of crowds*, pages 347–382. Springer, 2013.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/

9015-pytorch-an-imperative-style-high-performance-deep-learn
pdf.

[27] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[29] Vishwanath A Sindagi and Vishal M Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[30] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021.

[31] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[32] Ye Tian, Xiangxiang Chu, and Hongpeng Wang. Cctrans: Simplifying and improving crowd counting with transformer. *arXiv preprint arXiv:2109.14483*, 2021.

[33] Ibrahim Saygin Topkaya, Hakan Erdogan, and Fatih Porikli. Counting people by clustering person detector outputs. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 313–318. IEEE, 2014.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[36] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer. *arXiv preprint arXiv:2101.08461*, 2021.

[37] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5151–5159, 2017.

[38] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Weakly-supervised crowd counting learns from sorting rather than locations. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020.

[39] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 833–841, 2015.

[40] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016.