CounTr: An End-to-End Transformer Approach for Crowd Counting and Density Estimation

Haoyue Bai(✉), Hao He, Zhuoxuan Peng, Tianyuan Dai, and S.-H. Gary Chan

Hong Kong University of Science and Technology, Hong Kong, Hong Kong
{hbaiaa,gchan}@cse.ust.hk, {hheat,zpengac,tdaiaa}@connect.ust.hk

Abstract. Modeling context information is critical for crowd counting and desntiy estimation. Current prevailing fully-convolutional network (FCN) based crowd counting methods cannot effectively capture long-range dependencies with limited receptive fields. Although recent efforts on inserting dilated convolutions and attention modules have been taken to enlarge the receptive fields, the FCN architecture remains unchanged and retains the fundamental limitation on learning long-range relationships. To tackle the problem, we introduce CounTr, a novel end-to-end **tr**ansformer approach for crowd **coun**ting and density estimation, which enables capture global context in every layer of the Transformer. To be specific, CounTr is composed of a powerful transformer-based hierarchical encoder-decoder architecture. The transformer-based encoder is directly applied to sequences of image patches and outputs multi-scale features. The proposed hierarchical self-attention decoder fuses the features from different layers and aggregates both local and global context features representations. Experimental results show that CounTr achieves state-of-the-art performance on both person and vehicle crowd counting datasets. Particularly, we achieve the first position (159.8 MAE) in the highly crowded UCF_CC_50 benchmark and achieve new SOTA performance (2.0 MAE) in the super large and diverse FDST open dataset. This demonstrates CounTr's promising performance and practicality for real applications.

Keywords: Single image crowd counting · Transformer-based approach · Hierarchical architecture

1 Introduction

Crowd counting and density estimation has received increasing attention in computer vision, which is to estimate the number of objects (e.g., people, vehicle) in unconstrained congested scenes. The crowd scenes are often taken by a surveillance camera or drone sensor. Crowd counting enables a myriad of applications

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-25075-0_16.

L. Karlinsky et al. (Eds.): ECCV 2022 Workshops, LNCS 13806, pp. 207–222, 2023.
https://doi.org/10.1007/978-3-031-25075-0_16

Fig. 1. Visualization of attention maps: (a) Original image, (b) CNN-based methods, (c) Transformer-based methods, which extract global context information. Our observation is that a single layer of transformer can capture a larger range of context information than CNN-based methods.

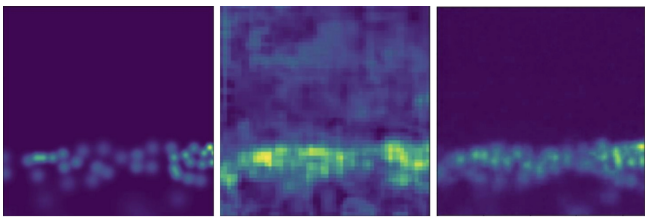

Fig. 2. Visualization of density maps: (a) Ground truth, (b) CNN-based methods, (c) Transformer-based methods. Our observation is that transformer-based methods achieves better visual quality on the generated density maps.

in the real world, such as public safety, video surveillance, and traffic management [14], [24], [26]. Benefiting from the rapid development of deep learning [13,25,29], fully convolutional network-based models have been the prevailing methods in crowd counting [39], [21,22].

Since extracting context feature representation is one of the major concerns in crowd estimation, building FCN-based models with multi-column architecture [12,34], dilated convolution [7,43] and attention mechanisms [10,28,44] has become a predominant design choice to enlarge the receptive fields and achieves significant advances in the field of crowd counting. However, the conventional fully convolutional network-based framework remains unchanged and retains the fundamental problem of convolutional neural network (CNN), which mainly focuses on small discriminate regions and cannot effectively capture the global context information. Estimating objects in crowded environments is still challenging to the community.

Recently, Transformer has achieved superior performance in image recognition [8,23], detection [6] and re-identification [11], due to its effective architecture on extracting the long-range context information. Thus, Transformer has the potential to address the aforementioned issues in crowd counting. The ability to model the long-range dependencies is suitable for better feature extraction and to make connections of target objects in crowded areas, as shown in

Fig. 1 and Fig. 2. This observation encourages us to build an end-to-end crowd counting model with a pure transformer. However, Transformer still needs to be specifically designed for crowd counting to tackle the unique challenges. (1) The object scales are varied and unevenly distributed in crowd images. Substantial efforts are needed to address this challenge by modeling both local and global context feature representations in the transformer-based crowd counting model. (2) Crowd counting not only relies on extracting strong multi-level features but also requires evaluating the generated density maps in terms of resolution and visual quality, which contains relative location information. Thus, a specifically designed decoder that effectively aggregates multi-level feature representations is essential for accurate crowd counting and high-quality density map generation.

In this work, we propose CounTr, a novel end-to-end transformer approach that can serve as a better substitute for FCN-based crowd counting methods, which is formed by a transformer-based hierarchical encoder-decoder architecture. The input image is split into fixed-size patches and is directly fed to the model with a linear embedding layer applied to obtain the feature embedding vectors for discriminative feature representation learning. In order to effectively capture contextual information and learn powerful representations, the transformer-based encoder is presented to enable multi-scale feature extraction. Secondly, we introduce a hierarchical self-attention decoder to effectively aggregates the extracted multi-level self-attention features. This self-attention decoder module is densely connected to the transformer-based encoder with skip connections. The whole framework can be integrated into an end-to-end learning paradigm.

Our main contributions can be summarized as follows:

(1) We propose CounTr, a novel end-to-end Transformer approach for single image crowd counting, which effectively captures the global context information and consistently outperforms the CNN-based baselines.
(2) We introduce a transformer-based encoder to enhance multi-scale and robust feature extraction, and we further propose a hierarchical self-attention decoder for better leveraging local and long-range relationships and generating high-quality density maps.
(3) Extensive experiments show the superiority of our proposed method. Our CounTr achieves new state-of-the-art performance on both person and vehicle crowd counting benchmarks.

2 Related Works

2.1 Crowd Counting and Density Estimation

Various CNN-based methods have been proposed over the years for single image crowd counting [2]. MCNN [46] is a pioneering work that utilizes the multi-column convolutional neural networks with different filter sizes to address the scale variation problem. Switching-CNN [32] introduces a patch-based switching module on the multi-column architecture to effectively enlarge the scale range. SANet [5] stacks several multi-column blocks with densely up-sample layers to generate

high-quality density maps. In order to enlarge the receptive fields, CSRNet [15] utilizes dilated convolutional operations and model larger context information. CAN [20] introduces a multi-column architecture that extracts features with multiple receptive fields and learns the importance of each feature at every image location to accommodate varied scales. SASNet [35] proposes a scale-adaptive selection network for automatically learning the internal correspondence between the scales and the feature levels. DSSINet [19] designs a deep structured scale integration network and a dilated multi-scale structural similarity loss for extracting structured feature representations and generating high-quality density maps. DENet [18] proposes a detection network and an encoder-decoder estimation network for accurately and efficiently counting crowds with varied densities. AMRNet [21] designs an adaptive mixture regression to effectively capture the context and multi-scale features from different convolutional layers and achieves more accurate counting performance. However, CNN-based approaches cannot effectively model the long-range dependencies, due to the fundamental problem of the limited receptive fields. Our CounTr is able to naturally model the global context and effectively extract multi-scale features with Transformers.

2.2 Transformers in Vision

Transformers [38] were first proposed for the sequence-to-sequence machine translation task. Recently, Transformers has achieved promising performance in the field of image classification, detection, and segmentation. Vision Transformer [8] directly applies to sequences of image patches for image classification and achieves excellent results compared to convolutional neural network-based baselines. Swin Transformer [23] introduces an accurate and efficient hierarchical Transformer with shifted windows to allow for cross-window connection. BEiT [4] introduces a self-supervised vision representation model, which learns bidirectional encoder representation from image transformers. DETR [6] utilizes a Transformer-based backbone and a set-based loss for object detection. SegFormer [41] unifies hierarchically structured Transformers with lightweight MLP decoders to build a simple, efficient yet powerful semantic segmentation framework. VoTr [27] introduces a voxel-based Transformer backbone to capture long-range relationships between voxels for 3D object detection. SwinIR [16] proposes a strong baseline model for image restoration based on the Swin Transformer. TransReID [11] proposes a pure transformer-based with jigsaw patch module to further enhance the robust feature learning for the object ReID framework. VisTR [40] presents a new Transformer-based video instance segmentation framework. The work in [36] introduces a token-attention module and a regression-token module to extract global context. Our CounTr extends the idea of Transformers on images and proposes an end-to-end method to apply Transformer to crowd counting. Compared with traditional vision transformers, CounTr benefits from the efficiency of capturing local and global context information via the transformer-based encoder and hierarchical self-attention decoder. CounTr achieves superior counting accuracy and is able to generate high-quality density maps.

3 Methodology

In this section, we present CounTr, a Transformer-based end-to-end crowd counting and density estimation framework. CounTr can extract multi-scale feature representation and enhance robustness through the transformer-based encoder and pixel shuffle operations [1]. We further propose a hierarchical self-attention decoder to facilitate the fusion of both local and long-range context information. The whole framework can be jointly trained in an end-to-end manner.

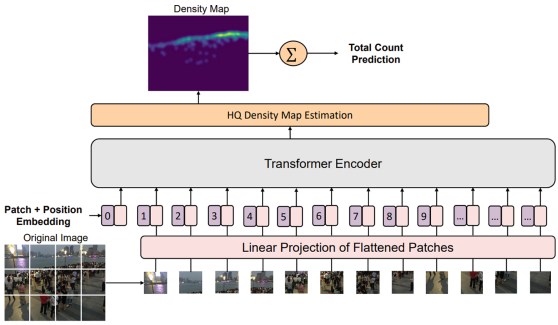

Fig. 3. A strong transformer-based crowd counting baseline. The input crowd image is split into fixed-size patches, linearly embedded, added with positional embeddings, fed to a standard Transformer encoder. The feature extracted by the Transformer encoder is rearranged and upsampled to the original input size, and the pixel value is summed up to predict the total counting number. The Transformer encoder for the image process was inspired by [8]

3.1 Preliminaries on Transformers for Crowd Counting

We introduce a transformer-based strong baseline for crowd counting and density estimation, as shown in Fig. 3. Our method has two main components: feature extraction, density estimation. We split and reshape the initial crowd image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$ (H, W and C are the height, width and the number of channel, respectively) into M fixed-sized flattened image patches$\{x_p^i | i = 1, 2, ..., M\}$. The standard Transformer encoder architecture consists of a multi-head self-attention module and a feed-forward network. All the transformer layers have a global receptive field, thus this addresses the CNN-based crowd counting methods' limited receptive field problem. The positional embedding is added with each image patch to provide position information, and the positional embedding is learnable. We linearly embed the patch sequences, add positional embedding, and feed to the standard transformer encoder.

The feature generated by the standard transformer encoder is rearranged and up-sampled to the original input size and generates high-quality density maps, which present the number of objects of each pixel. Finally, the total counting number is predicted by summing up all the pixel values within an image. We

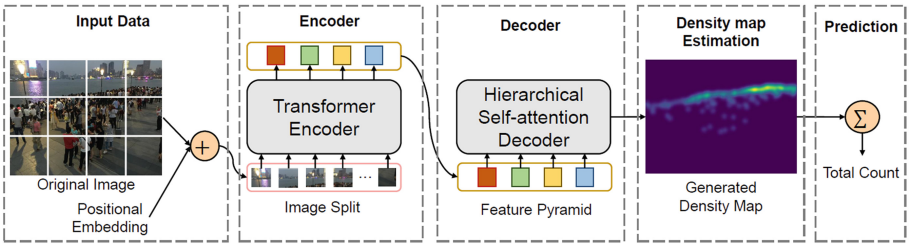

Fig. 4. The architecture of CounTr. CounTr split the original image into patches and added with a positional encoding before fed it into a transformer encoder. The encoder is a stack of swin transformer blocks [23] with different shifted windows and output feature pyramid embedding. We pass each output feature pyramid embedding of the encoder to a hierarchical self-attention decoder with a skip connection that predicts density map and total count number.

optimize the counting network by MSE loss and counting loss for global features. The MSE loss is the pixel-wise loss function, and the counting loss is the L1 constrain of the total object count. The widely used Euclidean loss is shown as follows: $L_E = \frac{1}{N} ||F(x_i;\theta) - y_i| |_2^2$, where θ indicates the model parameters, N means the number of pixels, x_i denotes the input image, and y_i is ground truth and $F(x_i;\theta)$ is the generated density map.

Though promising, pure transformer-based encoder has much less image-specific inductive bias than CNNs, traditional transformers (including ViT) have tokens with fixed scale. However, visual elements can be different in scale, whereas the object scales are diversified and the objects are usually unevenly distributed especially for crowd images. Thus, substantial efforts are needed to address this challenge by extracting both local and long-range features, and a specifically designed decoder is needed to effectively leverage multi-scale features for accurate crowd estimation.

3.2 The CounTr Model

We introduce the overall architecture of CounTr in Fig. 4, which consists of two main modules: shifted transformer-based encoder and hierarchical self-attention decoder. We split the original image into patches, embed the positional encoding, and fed the sequence into the shifted transformer encoder. The encoder is a stack of shifted transformer blocks with different shifted windows and patch shuffle operations. The generated feature pyramid embedding from the transformer encoder is fed into the hierarchical self-attention decoder by skip connections [31]. Finally, the density map estimation step predicts the density map and the total count number.

Shifted Transformer-Based Encoder. The Shifted transformer-based encoder aims to extract long-range contextual information and naturally enlarge the receptive field. CounTr incorporates swin transformer blocks [23] with patch

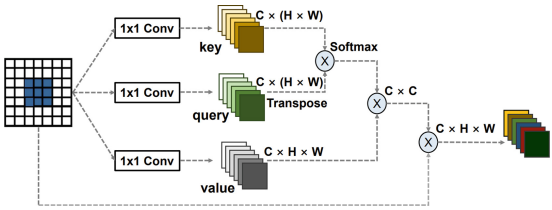

Fig. 5. The illustration of the global self-attention module. Global self-attention focuses on the whole space with masks on the channel-wise dimension.

shuffle operation as the encoder backbone. Using self-attention and encoder-decoder framework on this patches embeddings, the model globally captures all objects in a crowd scene using pair-wise relations between them, which can extract long-range context information and even use the whole image as context.

Standard transformer-based models split images into non-overlapping patches, losing local neighboring structures around the patches. We use shifted windows to generate patches with overlapping pixels. Patch shuffle operations further enhance local and long-range feature extraction. With the shift and shuffle operation, CounTr captures the local feature between short-range objects and increases the global discriminative capability of the local range object parts. In this way, CounTr can effectively capture local and long-range context feature representation.

The encoder part generates multi-level features. The final receptive field is 1/8 of the original resolution, and outputs multi-level feature pyramid embedding from different levels of the stacked shifted transformer blocks. The pyramid feature embedding includes both low-level (e.g., texture, color...) and high-level semantic representation, which can be used to facilitate the downstream hierarchical information congregation step.

Hierarchical Self-attention Decoder. The hierarchical self-attention module takes the pyramid feature embedding [17] as input, which leverages multi-level self-attention features to achieve accurate crowd estimation. It makes use of the feature pyramid to integrates both local and global relationships. We require short-range contextual information for neighboring pixels and also need long-range context information from the deeper layer of the transformer encoder with a large receptive field and high-level semantic information. Thus, our hierarchical self-attention module can cater to varied object scales with multi-level representation and effectively model the isolated small clusters in unconstrained crowd scenes.

The global self-attention mechanism in the hierarchical self-attention module further enhance autofocusing context information. This module consists of N layers with different level feature embedding as the input vector. we also add a separate convolution layer with filter size 1×1 at the beginning of each self-attention module to reduce computation complexity, which benefits reducing the computation consumption without sacrificing performance.

Table 1. Statistics of different datasets in our experiment. Min, Max and Avg denote the minimum, maximum, and average counting numbers per image, respectively.

Dataset	Year	Average resolution	Image number	Total	Min count	Max count	Avg count
UCF_CC_50 [12]	2013	2101 × 2888	50	63,974	94	4,543	1,280
SmartCity [45]	2018	1080×1920	50	369	1	14	7.4
Fudan-ShanghaiTech [9]	2019	1080 × 1920	15,000	394,081	9	57	27
Drone People [3]	2019	969 × 1482	3347	108,464	10	289	32.4
Drone Vehicle [3]	2019	991 × 1511	5303	198,984	10	349	37.5

Directly concatenate pyramid features in the decoder part may contain redundant information. Thus, we utilize the global self-attention module to capture short and long-range relationships, which calculate the weighted sum of values and assign weights to measure the importance of the multi-level pyramid features. Directly combining and up-sample operation only assigns the same weight for each input feature vector, an inappropriate level of features may have bad effects on the crowd estimation. Our hierarchical self-attention module eliminates the drawbacks of the fixed word token problem in the traditional vision transformer model and is suitable for adaptively varied scales.

The details are shown in Fig. 5. The global self-attention module first transfers input x to query Q_x, key K_x and value V_x:

$$Q_x = f(x), K_x = g(x), V_x = h(x). \tag{1}$$

The output weighted density map Y is computed by two kinds of matrix multiplications:

$$Y = \text{softmax}\left(Q_x K_x^T\right) V_x. \tag{2}$$

Our proposed hierarchical self-attention module can automatically focus on the most suitable feature scales and enlarge the receptive field with limited extra network parameters.

4 Experiments

In this section, we conduct numerical experiments to evaluate the effectiveness of our proposed CounTr. To provide a comprehensive comparison with baselines, we compare our proposed CounTr with SOTA algorithms on various crowd counting datasets.

4.1 Implementation Details and Datasets

We evaluate our CounTr on five challenging crowd counting datasets with different crowd levels: UCF_CC_50 [12], SmartCity [45], Fudan-ShanghaiTech [9],

Fig. 6. Typical examples of crowd counting data from different datasets. (a) SmartCity. (b) UCF_CC_50. (c) FDST. (d) Drone People. (e) Drone Vehicle.

Drone People [3] and Drone Vehicle [3]. As shown in Table 1, the statistics of the five datasets are listed with the information of publication year, image resolution, the number of dataset images, the total instance number of the datasets, the minimal count, the maximum count, and its average annotation number for the whole dataset. These datasets are commonly used in the field of crowd counting (see Fig. 6).

UCF_CC_50. [12] has 50 black and white crowd images and 63974 annotations, with the object counts ranging from 94 to 4543 and an average of 1280 persons per image. The original average resolution of the dataset is 2101×2888. This challenging dataset is crawled from the Internet. For experiments, UCF_CC_50 were divided into 5 subsets and we performed 5-fold cross-validation. This dataset is used to test the proposed method on highly crowded scenes.

SmartCity. [45] contains 50 images captured from 10 city scenes including sidewalk, shopping mall, office entrance, etc. This dataset consists of images from both outdoor and indoor scenes. As shown in Table 1, the average number of people in SmartCity is 7.4 per image. The maximum count is 14 and the minimum count is 1. This dataset can be used to test the generalization ability of crowd counting methods on very sparse crowd scenes.

Fudan-ShanghaiTech. [9] is a large-scale crowd counting dataset, which contains 100 videos captured from 13 different scenes. FDST includes 150,000 frames and 394,081 annotated heads, which is larger than previous video crowd counting datasets in terms of frames. The training set of the FDST dataset consists

of 60 videos, 9000 frames, and the testing set contains the remaining 40 videos, 6000 frames. Some examples are shown in Fig. 6. The maximum count number is 57, and the minimum count number is 9. The number of frames per second (FPS) for FDST is 30. The statistics of FDST is shown in 1

Table 2. Performance comparison on UCF_CC_50 dataset.

Algorithm	UCF_CC_50	
	MAE	MSE
MCNN [46]	377.6	509.1
CP-CNN [33]	298.8	320.9
ConvLSTM [42]	284.5	297.1
CSRNet [15]	266.1	397.5
DSSINet [19]	216.9	302.4
CAN [20]	212.2	243.7
PaDNet [37]	185.8	278.3
SASNet [35]	161.4	234.5
CounTr (ours)	**159.8**	**173.3**

Table 3. Performance comparison on SmartCity dataset.

Algorithm	SmartCity	
	MAE	MSE
SaCNN [45]	8.60	11.60
YOLO9000 [30]	3.50	4.70
MCNN [46]	3.47	3.78
CSRNet [15]	3.38	3.89
DSSINet [19]	3.50	4.32
AMRNet [21]	3.87	4.91
DENet [18]	3.73	4.21
CounTr (ours)	**3.09**	**3.63**

Drone People. [3] is modified from the original VisDrone2019 object detection dataset with bounding boxes of targets to crowd counting annotations. The original category pedestrian and people are combined into one dataset for people counting. The new people annotation location is the head point of the original people bounding box. This dataset consists of 2392 training samples, 329 validation samples, and 626 test samples. Some examples are shown in Fig. 6. The average count number for the drone people dataset is 32.4 per image (Table 3).

For crowd counting, two metrics are used for evaluation, Mean Absolute Error (MAE) and Mean Squared Error (MSE), which are defined as: MAE $= \frac{1}{N}\sum_{i=1}^{N}|C_i - \widehat{C}_i|$, MSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}|C_i - \widehat{C}_i|^2}$, where N is the total number of test images, C_i means the ground truth count of the i-th image, and $\widehat{C}_i$ represents the estimated count. To evaluate the visual quality of the generated density maps, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity in Images (SSIM) are often used [33].

We adopt the geometry-adaptive kernels to address the highly congested scenes. We follow the same method of generating density maps in [46], i.e., the ground truth is generated by blurring each head annotation with a Gaussian kernel. The geometry-adaptive kernel is defined as follows:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) \times G_{\sigma_i}(x), \text{with } \sigma_i = \beta \bar{d}_i, \tag{3}$$

where x denotes the pixel position in an image. For each target object, x_i in the ground truth, which is presented with a delta function $\delta(x - x_i)$. The ground

Table 4. Performance comparison on the Fudan-ShanghaiTech dataset.

Algorithm	FDST	
	MAE	MSE
MCNN [46]	3.77	4.88
ConvLSTM [42]	4.48	5.82
LSTN [9]	3.35	4.45
DENet [18]	2.26	3.29
CounTr (ours)	**2.00**	**2.50**

Table 5. Performance comparison on drone-based datasets.

Algorithm	Drone people		Drone vehicle	
	MAE	MSE	MAE	MSE
MCNN [46]	16.4	39.1	14.9	21.6
CSRNet [15]	12.1	36.7	10.9	16.6
SACANet [3]	10.5	35.1	8.6	12.9
DSSINet [19]	13.4	19.3	10.3	15.5
AMRNet [21]	14.3	22.2	10.5	14.7
DENet [18]	10.3	16.1	6.0	10.3
CounTr (ours)	**7.6**	**12.5**	**5.1**	**8.9**

truth density map $F(x)$ is generated by convolving $\delta(x - x_i)$ with a normalized Gaussian kernel based on parameter σ_i. And $\bar{d}_i$ shows the average distance of the k nearest neighbors.

We compare our proposed CounTr with the state-of-the-art crowd counting algorithms, including MCNN [46], CSRNet [15], SACANet [3], CAN [20], PaDNet [37], SASNet [35], DSSINet [19], DENet [18], and AMRNet [21], etc. In our implementation, the input image is batched together, we apply 0-padding adequately to ensure that they all have the same dimensions. Our framework was implemented with PyTorch 1.7.1 and CUDA 11.3. We conducted experiments on GeForce RTX 3090. More implementation details can be found in the Appendix.

4.2 Results and Discussion

In this section, we evaluate and analyze the results of CounTr on five datasets: FDST, UCF_CC_50, SmartCity, Drone People, and Vehicle. These datasets are taken by a surveillance camera or drone sensor, which represents a different level of scale variation and isolated small clusters [3].

Illustrative Results on UCF_CC_50 Dataset. As shown in Table 2, CounTr achieves the best MAE and MSE performance on UCF_CC_50, followed by CNN-based methods, such as MCNN, DSSINet, etc. The traditional convolutional

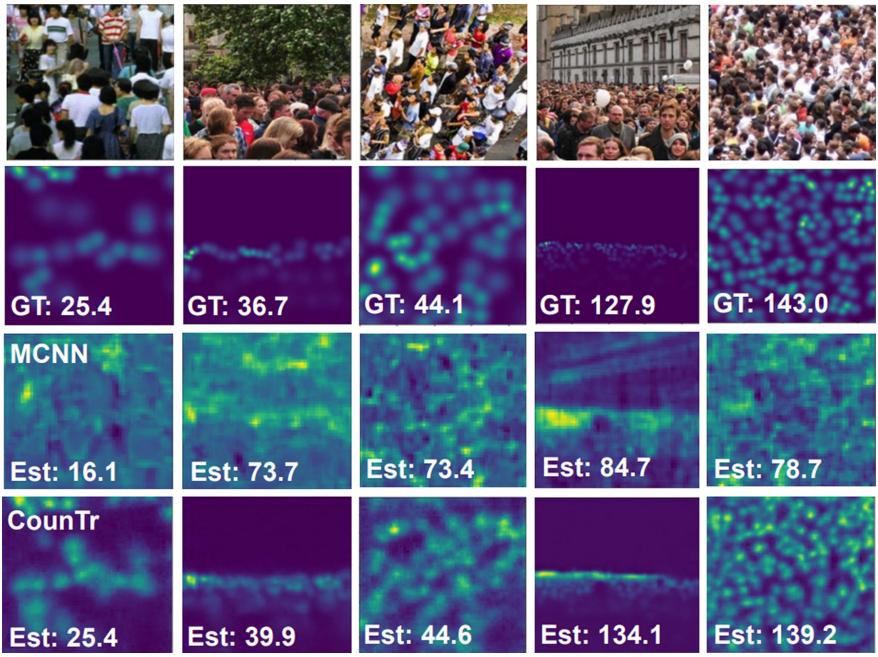

Fig. 7. Visualization of the generated density maps. The first row shows the original image, the second row presents ground truth density maps, the third row visualizes the density maps generated by CNN-based method MCNN [46]. The last row shows the density maps of our proposed CounTr. (Better viewed in the zoom-in mode)

crowd counting approaches are constrained by the limited receptive field in the CNN-based backbone. This dataset is highly crowded, and CounTr further improves the performance in UCF_CC_50 by capturing both local and global context information, which is crucial to address the large-scale variations in surveillance-based crowd datasets.

Illustrative Results on SmartCity Dataset. The results for the SmartCity dataset are shown in Table 4. We can see that the proposed CounTr framework achieves the SOTA performance compared with the various crowd counting baselines. The superior performance of CounTr confirms the possibility of improving the crowd estimation accuracy via better extracting the multi-scale feature representations and enlarge the receptive field by introducing transformer-based architecture into the crowd counting task.

Illustrative Results on the Fudan-Shanghaitech Dataset. CounTr achieves the state-of-the-art performance following the convolutional baselines such as MCNN and DSSINet. The results are shown in Table 4. Notice that CounTr achieves better performance, even compared with advanced ConvLSTM and LSTN, which incorporate extra temporal information. This may be because the LSTM-based methods cannot effectively extract the context information

Table 6. The visual quality comparison on different datasets. The baselines are implemented by ourselves.

Algorithm	Drone people		Drone vehicle	
	PSNR	SSIM	PSNR	SSIM
AMRNet [21]	16.40	0.31	18.50	0.62
DSSINet [19]	30.10	0.96	34.30	0.98
DENet [18]	35.80	0.98	36.00	0.99
CounTr (ours)	**38.03**	**0.99**	**37.14**	**0.99**

Table 7. Compared with other transformer-based methods on drone people. baselines are implemented by ourselves.

Algorithm	MAE	MSE
Vision Transformer [8]	8.9	18.4
Swin Transformer [23]	8.5	17.9
CounTr (ours)	**7.6**	**12.5**

and introduce redundant information. The superior performance further demonstrates the effectiveness of CounTr.

Illustrative Results on the Drone People Dataset. We also compare our CounTr with the different crowd counting algorithms on the Drone People dataset. We observe that our method achieves SOTA performance. The detailed experimental results for Drone People are shown in Table 5. From Table 5, the proposed CounTr method achieves much better counting accuracy than other CNN-based methods in terms of MAE (7.6) and MSE (12.5) for Drone People. This demonstrates the superiority of CounTr.

Illustrative Results on Drone Vehicle Dataset. To test the generalization ability of CounTr on other object counting tasks except for people counting, we compare CounTr with the different crowd counting methods on the Drone Vehicle dataset. The results are shown in Table 5. Our method consistently achieves good performance in terms of MAE and MSE with the non-trivial improvement compared with other counting methods. Specifically, CounTr achieves 5.1 MAE and 8.9 MSE, which is much better than previous crowd counting algorithms, such as MCNN (14.9 MAE) and DSSINet (10.3 MAE). This demonstrates the superiority of CounTr and its potential to be practically useful.

4.3 Ablation Study

In this section, we first compare CounTr with advanced transformer-based methods, such as ViT [8], and Swin Transformer [23]. This is to test whether directly applying Transformer-based methods to crowd counting tasks can improve the counting accuracy. We conduct experiments on the Drone People dataset. The

Table 8. Ablation study on drone people.

Algorithm	MAE	MSE
CounTr w/o pixel shuffle	8.17	15.31
CounTr w/o self-attention	8.15	14.11
CounTr (ours)	**7.60**	**12.50**

detailed results are shown in Table 7. We observe that naively using ViT can achieve only 8.9 MAE average accuracy, significantly lower than our CounTr method. This may be due to the lack of capturing local context features between the non-overlapping patches, and inappropriate decoder layers. This confirms that the hierarchical self-attention decoder is needed for accurate crowd estimation.

As shown in Table 8, the results of CounTr without pixel shuffle operations are 8.17 MAE and 15.31 MSE, which is lower than our final framework CounTr. We also conduct an ablation study on the self-attention module. The accuracy without self-attention is 8.15 MAE, which is much lower than our proposed CounTr (7.6 MAE). This also shows the effectiveness of the self-attention module and pixel shuffle operations to facilitate accurate crowd estimation.

4.4 Visualization on Density Maps

As shown in Fig. 7, we also visualize the crowd images and their corresponding density maps with different crowd levels. The first row is the original images, the second row is the corresponding ground truth, and the last row is the generated density maps. Besides, the visual quality results in terms of PSNR and SSIM are shown in Table 6. It can be seen that CounTr achieves consistently better performance on various crowd counting datasets. Both the qualitative and the quantitative results demonstrate the effectiveness of our proposed method to generate high-quality density maps.

5 Conclusion

In this paper, we propose CounTr, a novel Transformer-based end-to-end framework for crowd counting and density estimation. CounTr consists of two main modules: a shifted transformer-based encoder and a hierarchical self-attention decoder for better capturing short and long-range context information. Experimental results show that the final CounTr framework outperforms the CNN-based baselines and achieves new state-of-the-art performance on both person and vehicle crowd counting datasets.

References

1. Aitken, A., Ledig, C., Theis, L., Caballero, J., Wang, Z., Shi, W.: Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. arXiv:1707.02937 (2017)
2. Bai, H., Mao, J., Chan, S.H.G.: A survey on deep learning-based single image crowd counting: Network design, loss function and supervisory signal. Neurocomputing (2022)
3. Bai, H., Wen, S., Gary Chan, S.H.: Crowd counting on images with scale variation and isolated clusters. In: ICCVW (2019)
4. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv:2106.08254 (2021)
5. Cao, X., Wang, Z., Zhao, Y., Su, F.: Scale aggregation network for accurate and efficient crowd counting. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11209, pp. 757–773. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01228-1_45
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
7. Deb, D., Ventura, J.: An aggregated multicolumn dilated convolution network for perspective-free counting. In: CVPR Workshops (2018)
8. Dosovitskiy, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
9. Fang, Y., Zhan, B., Cai, W., Gao, S., Hu, B.: Locality-constrained spatial transformer network for video crowd counting. In: ICME (2019)
10. Gao, J., Wang, Q., Yuan, Y.: Scar: Spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing (2019)
11. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. arXiv:2102.04378 (2021)
12. Idrees, H., Saleemi, I., Seibert, C., Shah, M.: Multi-source multi-scale counting in extremely dense crowd images. In: CVPR (2013)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
14. Lempitsky, V., Zisserman, A.: Learning to count objects in images. In: NeurIPS (2010)
15. Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: CVPR (2018)
16. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. arXiv:2108.10257 (2021)
17. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
18. Liu, L., et al.: Denet: A universal network for counting crowd with varying densities and scales. In: TMM (2020)
19. Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., Lin, L.: Crowd counting with deep structured scale integration network. In: ICCV (2019)
20. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: CVPR (2019)
21. Liu, X., Yang, J., Ding, W., Wang, T., Wang, Z., Xiong, J.: Adaptive mixture regression network with local counting map for crowd counting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12369, pp. 241–257. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_15

22. Liu, Y., et al.: Crowd counting via cross-stage refinement networks. In: TIP (2020)
23. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv:2103.14030 (2021)
24. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: ICCV (2019)
25. Mao, J., et al.: One million scenes for autonomous driving: Once dataset. arXiv preprint arXiv:2106.11037 (2021)
26. Mao, J., Shi, S., Wang, X., Li, H.: 3d object detection for autonomous driving: A review and new outlooks. arXiv preprint arXiv:2206.09474 (2022)
27. Mao, J., et al.: Voxel transformer for 3d object detection. In: ICCV (2021)
28. Miao, Y., Lin, Z., Ding, G., Han, J.: Shallow feature based dense attention network for crowd counting. In: AAAI (2020)
29. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016)
30. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: CVPR (2017)
31. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
32. Sam, D.B., Surya, S., Babu, R.V.: Switching convolutional neural network for crowd counting. In: CVPR (2017)
33. Sindagi, V.A., Patel, V.M.: Generating high-quality crowd density maps using contextual pyramid cnns. In: ICCV (2017)
34. Sindagi, V.A., Patel, V.M.: Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: ICCV (2019)
35. Song, Q., et al.: To choose or to fuse? scale selection for crowd counting. In: AAAI (2021)
36. Sun, G., Liu, Y., Probst, T., Paudel, D.P., Popovic, N., Van Gool, L.: Boosting crowd counting with transformers. arXiv:2105.10926 (2021)
37. Tian, Y., Lei, Y., Zhang, J., Wang, J.Z.: Padnet: Pan-density crowd counting. In: TIP (2019)
38. Vaswani, A., et al.: Attention is all you need. In: NeurIPS (2017)
39. Wang, B., Liu, H., Samaras, D., Hoai, M.: Distribution matching for crowd counting. arXiv:2009.13077 (2020)
40. Wang, Y., et al.: End-to-end video instance segmentation with transformers. In: CVPR (2021)
41. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. arXiv:2105.15203 (2021)
42. Xiong, F., Shi, X., Yeung, D.Y.: Spatiotemporal modeling for crowd counting in videos. In: ICCV (2017)
43. Yan, Z., Zhang, R., Zhang, H., Zhang, Q., Zuo, W.: Crowd counting via perspective-guided fractional-dilation convolution. In: TMM (2021)
44. Zhang, A., et al.: Relational attention network for crowd counting. In: ICCV (2019)
45. Zhang, L., Shi, M., Chen, Q.: Crowd counting via scale-adaptive convolutional neural network. In: WACV (2018)
46. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: CVPR (2016)