Journal of King Saud University - Computer and Information Sciences 36 (2024) 101972

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Full Length Article

Cross-scale Vision Transformer for crowd localization

Shuang Liu [a], Yu Lian [a], Zhong Zhang [a,*], Baihua Xiao [b], Tariq S. Durrani [c]

[a] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin, 300387, China
[b] The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[c] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow Scotland, UK

ARTICLE INFO

Keywords:
Crowd localization
Multi-scale information fusion
Long-range context dependencies
Adaptive windows

ABSTRACT

Crowd localization can provide the positions of individuals and the total number of people, which has great application value for security monitoring and public management, meanwhile it meets the challenges of lighting, occlusion and perspective effect. In recent times, Transformer has been applied in crowd localization to overcome these challenges. Yet such kind of methods only consider to integrate the multi-scale information once, which results in incomplete multi-scale information fusion. In this paper, we propose a novel Transformer network named Cross-scale Vision Transformer (CsViT) for crowd localization, which simultaneously fuses multi-scale information during both the encoder and decoder stages and meanwhile building the long-range context dependencies on the combined feature maps. To this end, we design the multi-scale encoder to fuse the feature maps of multiple scales at corresponding positions so as to obtain the combined feature maps, and meanwhile design the multi-scale decoder to integrate the tokens at multiple scales when modeling the long-range context dependencies. Furthermore, we propose Multi-scale SSIM (MsSSIM) loss to adaptively compute head regions and optimize the similarity at multiple scales. Specifically, we set the adaptive windows with different scales for each head and compute the loss values within these windows so as to enhance the accuracy of the predicted distance transform map. We perform comprehensive experiments on five public datasets, and the results obtained validate the effectiveness of our method.

1. Introduction

In recent times, deep learning has been applied and developed across diverse research domains (Chen et al., 2023; Si et al., 2023; He et al., 2022; Liu et al., 2023). Crowd analysis based on deep learning, including identification, counting and localization of pedestrians, has attracted great interest among researchers due to its wide applications in various domains such as intelligence monitoring and public safety (Lin et al., 2023; Basalamah et al., 2023; Zhao and Li, 2023; Gong et al., 2023). Crowd localization as an important task of crowd analysis focuses on predicting the position of each individual and estimate the total number of human heads in the crowd (Song et al., 2021; Wang et al., 2023a; Liang et al., 2022b; Abousamra et al., 2021). Compared with crowd counting only estimating the total number of individuals (Zhang et al., 2016; Wang et al., 2023c; Qiu et al., 2017; Wang et al., 2023b), crowd localization provides detailed information of the crowd spatial distribution, which could provide efficient crowd management and emergency response.

Crowd localization faces significant challenges, such as illumination, occlusion and perspective effect. Hence, many approaches are proposed to overcome these challenges (Abousamra et al., 2021; Zhang et al., 2016; Liu et al., 2019c; Ma et al., 2019). The approaches used mainly fall into the following three categories.

Firstly, the detection-based approaches (Liu et al., 2019c; Sam et al., 2021; Lian et al., 2019; Wang et al., 2021c) typically rely on the nearest neighbor distances between head points to generate pseudo ground-truth bounding boxes which are treated as the training supervision information. Secondly, the regression-based approaches (Song et al., 2021; Liang et al., 2022b) directly regress the head coordinates without generating pseudo ground-truth bounding boxes, and output the corresponding confidence scores. Then, the confidence scores are used

* Corresponding author.
E-mail address: zhong.zhang8848@gmail.com (Z. Zhang).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2024.101972
Received 30 December 2023; Received in revised form 4 February 2024; Accepted 13 February 2024
Available online 15 February 2024

Fig. 1. (a) and (c) indicate that the head sizes in the front and back vary significantly. The corresponding point-level annotations are shown in (b) and (d), with green markers indicating the ground-truth heads.

to determine the final head coordinates. Nevertheless, these regression-based approaches neglect the correlation between the head point and its neighboring pixels, causing suboptimal localization performance. Thirdly, the map-based approaches (Abousamra et al., 2021; Liang et al., 2022c; Idrees et al., 2018; Xu et al., 2022) generate the training maps based on the annotation points and their neighbors where the pixels in the map have large values if they are close to the head point. These maps contain the head positions and rich spatial information, which are treated as supervision information in the training phase to produce accurate localization performance.

Recently, Transformer has gained significant popularity in the computer vision tasks (Zhang et al., 2023; Liu et al., 2021b; Touvron et al., 2021; Wu et al., 2021). Some methods (Liang et al., 2022b; Deng et al., 2023) utilize Transformer to construct long-range context dependencies to enhance the representation capacity of the network for crowd images. Fig. 1 appears multi-scale heads in the crowds, that is the head sizes varies greatly which is caused by the perspective effect of the camera, and this phenomenon degrades the performance of crowd localization. However, the existing Transformer-based approaches (Liang et al., 2022b; Deng et al., 2023) only consider to integrate the multi-scale information once, which results in incomplete multi-scale information fusion.

In this paper, we propose a novel Transformer network named Cross-scale Vision Transformer (CsViT) for crowd localization, which simultaneously fuses multi-scale information in the encoding and decoding stages and meanwhile builds the long-range context dependencies on the combined feature maps. Specifically, we combine the feature maps of each scale with the feature maps of shallower scales at the corresponding positions in the encoding stage, and obtain the combined feature maps. Since the receptive field sizes of feature maps in the shallow and deep layers are different, integrating the feature maps across multiple scales can effectively deal with the issue of multi-scale heads. To capture long-range context dependencies, we construct global attention information on the combined feature maps through Cross-Shaped Window (CSWin) Transformer. In the decoding stage, we combine the outputs of CSWin Transformer layer by layer, and therefore the predicted distance transform map also contains the multi-scale information. In a word, the deep model focuses on fusing the same head at multiple scales in the encoding stage, and meanwhile it pays attention to fusing long-range context dependencies between heads at multiple scales in the decoding stage. Hence, the two-stage multi-scale information fusion improves the representation capability of the deep model.

Recently, some methods apply SSIM-based loss (Liang et al., 2022c; Cao et al., 2018) to optimize the deep model in the field of crowd analysis. Nevertheless, these methods solely compare the similarity within a single-scale region, ignoring the fact that the region occupied by each head is different. Hence, we propose a novel loss named Multi-scale SSIM (MsSSIM) loss which could adaptively compute head regions and optimize the similarity at multiple scales. Specifically, we select the regions at different scales for each head by considering the distance between the heads, and then integrate the losses of multi-scale regions, which could optimize the accuracy of the predicted distance transform map.

In summary, our contributions mainly lie in three folds:

(1) We propose CsViT for crowd localization, which simultaneously fuses multi-scale information in the encoding and decoding stages when building long-range context dependencies. The representation capability of the deep model is enhanced by the two-stage multi-scale information fusion strategy.

(2) We propose the MsSSIM loss to optimize the multi-scale information fusion by comparing the similarity of different scales head regions. It could further reduce the negative responses caused by the background and improve the accuracy of the predicted distance transform map.

(3) We evaluate the proposed method on five public datasets, i.e., ShanghaiTech, UCF-QNRF, JHU-Crowd++, UCF_CC_50 and NWPU-Crowd, and the results indicate that our method achieves the state-of-the-art performance in both crowd counting and localization tasks.

2. Related work

2.1. Detection-based approaches

Most detection-based approaches (Sam et al., 2021; Wang et al., 2021c; Liu et al., 2018) generate pseudo ground-truth bounding boxes based on dot labels or manually label part of the bounding boxes as supervision information. Liu et al. (2019c) initialize the size of the bounding box with the nearest neighbor distance between the heads, and an iterative update method is used to adjust the bounding boxes. Considering that the correlation between the size of the human head and the distance from the head to the camera, Lian et al. (2019) predict the size of the bounding boxes by considering the assistance of depth information. Nevertheless, these approaches show suboptimal performance in terms of accurate localization in extremely dense scenes due to occlusion and blurring.

2.2. Regression-based approaches

Regression-based approaches (Song et al., 2021; Liang et al., 2022b) can regress the coordinates of head points and output the confidence scores. Song et al. (2021) present a framework for counting and localization based on points, where a series of point proposals are regressed to indicate the heads using predefined anchor points. Liang et al. (2022b) propose a crowd localization model based on DETR (Carion et al., 2020), which applies trainable query instances instead of a large number of predefined anchor points. Nevertheless, these approaches lack the correlation information between the head point and surrounding pixels, so the localization performance is not accurate enough.

2.3. Map-based approaches

Map-based approaches (Wang et al., 2023a; Abousamra et al., 2021; Liang et al., 2022c) generate the training maps to guide the model training, which can reflect the relationship between the head points and the neighboring pixels. Idrees et al. (2018) conduct the head localization using the density map, where the head point location is determined by identifying the maximum value within the local area of the density map. Abousamra et al. (2021) present a topological approach which slightly dilates the head points into a dot mask, and

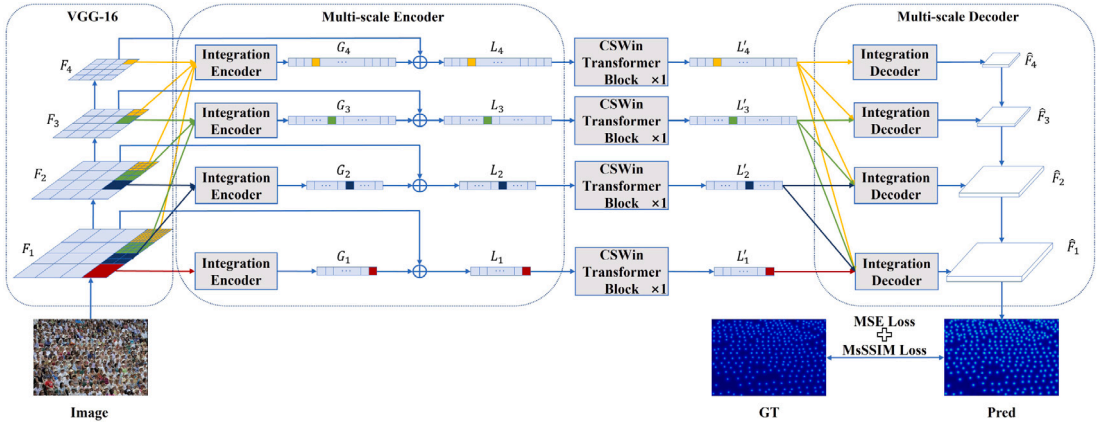

Fig. 2. The overall architecture of CsViT. It mainly consists of the VGG-16 backbone network, multi-scale encoder, CSWin Transformer Block and multi-scale decoder.

they utilize the dot mask map as supervision information. Xu et al. (2022) generate the distance label maps based on the distance between the head points, which could avoid the issue of overlapping heads in the dense regions. Liang et al. (2022c) propose the Focal Inverse Distance Transform (FIDT) map which could well represent the correlation between the head points and other pixels. Wang et al. (2023a) generate robust binary maps as the ground-truth through an adaptive threshold map for crowd localization. These approaches take full advantage of spatial information, but they ignore completed multi-scale information fusion.

2.4. Vision transformer

Transformer (Vaswani et al., 2017) is first proposed in natural language processing to capture long-range context dependencies for text content. Dosovitskiy et al. (2021) propose Vision Transformer (ViT) in computer vision, which achieves competitive performance compared to CNN. However, global self-attention in Transformer is computationally intensive. Hence, Liu et al. (2021b) present the Swin Transformer, which employs the shifted windows to perform self-attention. Afterwards, Dong et al. (2022) present CSWin Transformer to perform long-range context dependencies by using the cross-shaped window. These methods (Liu et al., 2021b; Dong et al., 2022) greatly reduce the computation amount of the self-attention.

In crowd analysis, some approaches (Liang et al., 2022b; Deng et al., 2023; Liang et al., 2022a, 2023) implement crowd counting and localization using Transformer. Liang et al. (2022b) present an end-to-end Transformer model for crowd counting and localization, which utilizes query instances to directly return the coordinates of head points in the image. Liang et al. (2023) first present to perform crowd counting using visual language knowledge, and design a multi-modal ranking loss to guide the learning of ViT by constructing ranking text prompts for unsupervised crowd counting. However, these approaches do not consider completed multi-scale information fusion during the learning process. Instead, our method fuses the multi-scale information of heads in the encoding stage, the decoding stage and the loss function, simultaneously.

3. Approach

3.1. Overview of CsViT

Fig. 2 depicts the framework of the proposed CsViT. Firstly, we utilize VGG-16 (Simonyan and Zisserman, 2015) as the backbone to extract the feature maps of four scales for each image. In the multi-scale encoding stage, the feature maps of each scale are combined with the feature maps of the shallower scales using the integration encoder, and then they are flattened into the tokens. After a residual connection, we apply the CSWin Transformer Block to learn the long-range context dependencies. In the multi-scale decoding stage, the outputs of CSWin Transformer Blocks are combined by the integration decoder. Finally, we utilize MSE loss and the proposed MsSSIM loss to optimize the network.

3.2. Multi-scale encoder

The head sizes in the crowd vary greatly, and this appears the multi-scale phenomenon. Hence, we design the multi-scale encoding to achieve completed multi-scale information fusion by combining different scales feature maps at the corresponding positions.

For a given image $I \in \mathbb{R}^{C\times H\times W}$, we obtain the feature maps with four scales $F_i \in \mathbb{R}^{C_i\times\frac{H}{2^i}\times\frac{W}{2^i}}$ $(i = 1, 2, 3, 4)$ using VGG-16 (Simonyan and Zisserman, 2015), where C, H and W are the number of channels, height and width respectively. In the multi-scale encoder, we utilize the integration encoder to combine the feature maps of each scale with the feature maps of the shallower scales at the corresponding positions. We take the deepest feature maps $F_4 \in \mathbb{R}^{C_4\times\frac{H}{2^4}\times\frac{W}{2^4}}$ as an example to illustrate the integration encoder, as depicted in Fig. 3. The region with the size of 1×1 in F_4 corresponds to the regions with the size of $2^{4-i} \times 2^{4-i}$ in the shallow feature maps F_i $(i = 1, 2, 3)$. We utilize the integration encoder to combine the corresponding regions of the feature maps. Specifically, the feature maps F_4 are flattened into the tokens $D_4 \in \mathbb{R}^{C_4\times\left(\frac{H}{2^4}\times\frac{W}{2^4}\right)}$, and the shallow feature maps are flattened to the tokens $D_i \in \mathbb{R}^{(C_i\times N_i)\times\left(\frac{H}{2^4}\times\frac{W}{2^4}\right)}$, where $N_i = 2^{4-i} \times 2^{4-i}$. Then, we map these tokens D_i to $D'_i \in \mathbb{R}^{C_4\times\left(\frac{H}{2^4}\times\frac{W}{2^4}\right)}$ $(i = 1, 2, 3, 4)$ by the linear layers. Afterwards, we obtain $G_4 \in \mathbb{R}^{C_4\times\left(\frac{H}{2^4}\times\frac{W}{2^4}\right)}$ by summing D'_i $(i = 1, 2, 3, 4)$. Similarly, we can obtain $G_i \in \mathbb{R}^{C_i\times\left(\frac{H}{2^i}\times\frac{W}{2^i}\right)}$ $(i = 1, 2, 3)$ using the integration encoder. Finally, we generate $L_i \in \mathbb{R}^{C_i\times\left(\frac{H}{2^i}\times\frac{W}{2^i}\right)}$ $(i = 1, 2, 3, 4)$ via a residual connection as the input of CSWin Transformer Block.

The feature maps at different scales possess different receptive field sizes, and therefore our multi-scale encoder could effectively utilize the information of feature maps of each scale by combining them across scales, which is beneficial to localizing the head positions at different scales.

3.3. CSWin Transformer Block

We apply the CSWin Transformer Block (Dong et al., 2022) to learn long-range context dependencies for each scale. Specifically, for the input tokens $L_i \in \mathbb{R}^{C_i\times\left(\frac{H}{2^i}\times\frac{W}{2^i}\right)}$, we perform the cross-shaped window (CSWin) self-attention in the CSWin Transformer Block with the multi-head strategy. Each CSWin self-attention is designed with K heads,

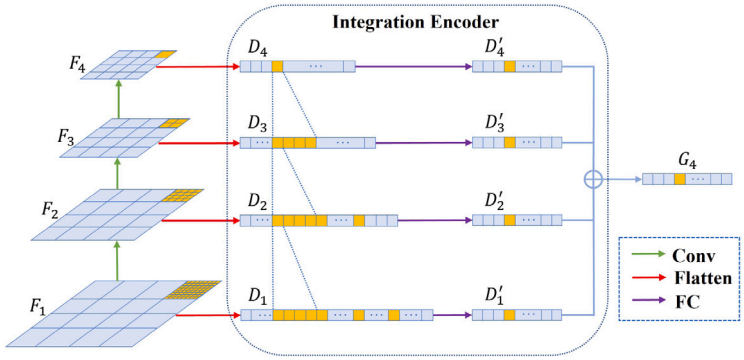

Fig. 3. The structure of the integration encoder. F_1, F_2, F_3 and F_4 are the feature maps with four scales extracted from VGG-16, and they are combined in the corresponding regions to obtain G_4.

and these heads are evenly divided into two groups, half of which perform horizontal stripes self-attention and the other half perform vertical stripes self-attention.

As for the horizontal stripes self-attention, we evenly divide L_i into J horizontal stripes $\left[L_i^1, \ldots, L_i^j, \ldots, L_i^J\right]$, where $L_i^j \in \mathbb{R}^{C_i \times \left(\frac{H}{J \times 2^i} \times \frac{W}{2^i}\right)}$. The horizontal stripes self-attention with the kth head is formulated as:

$$H_{Att-k}\left(L_i\right) = \left[Y_{ik}^1, \ldots, Y_{ik}^j, \ldots, Y_{ik}^J\right] \tag{1}$$

$$Y_{ik}^j = softmax\left(\frac{Q_{ik}^j \left(K_{ik}^j\right)^T}{\sqrt{d_k}}\right) V_{ik}^j \tag{2}$$

where $Q_{ik}^j = \left(L_i^j\right)^T W_{ik}^Q$, $K_{ik}^j = \left(L_i^j\right)^T W_{ik}^K$ and $V_{ik}^j = \left(L_i^j\right)^T W_{ik}^V$ are the query, key and value of the kth head respectively, and d_k is the dimension of the kth head. Here, W_{ik}^Q, W_{ik}^K and $W_{ik}^V \in \mathbb{R}^{C_i \times d_k}$ represent the projection matrices of the kth head, respectively. In the same way, the kth head of the vertical stripes self-attention is denoted as $V_{Att-k}(L_i)$. The CSWin self-attention is defined by concatenating the two parts:

$$CSWin = Concat(head_1, \ldots, head_K) W^O \tag{3}$$

$$head_k = \begin{cases} H_{Att-k}(L_i) & \text{if } k = 1, \ldots, \frac{K}{2} \\ V_{Att-k}(L_i) & \text{if } k = \frac{K}{2} + 1, \ldots, K \end{cases} \tag{4}$$

where $W^O \in \mathbb{R}^{C_i \times C_i}$ is the projection matrix.

Finally, the CSWin Transformer Block is formulated as:

$$L_i' = MLP\left(LN(S_i)\right) + S_i \tag{5}$$

$$S_i = CSWin\left(LN(L_i)\right) + L_i \tag{6}$$

where $L_i' \in \mathbb{R}^{C_i \times \left(\frac{H}{2^i} \times \frac{W}{2^i}\right)}$ is the output of the CSWin Transformer Block, LN is the layer normalization and MLP is the multi-layer perceptron.

3.4. Multi-scale decoder

The tokens with different scales L_i' outputted by the CSWin Transformer Blocks contain rich scale information, and therefore we apply the multi-scale combination strategy to decode them so as to fully fuse the multi-scale information.

In the decoding stage, we utilize the integration decoder to combine the tokens of each scale with the tokens of deeper scales. We take the shallowest feature maps $\hat{F}_1$ as an example:

$$\hat{F}_1 = Conv_{3\times3}\left(ID\left(L_1', L_2', L_3', L_4'\right) + UP\left(\hat{F}_2\right)\right) \tag{7}$$

where $Conv_{3\times3}$, ID and UP denote the convolution operation using the kernel size of 3 × 3, the integration decoder and the upsampling operation, respectively. Here, for the integration decoder, we first reshape the tokens L_i' into the feature maps with the same size of F_i $(i = 1, 2, 3, 4)$, and then resize them into the same size of F_1. Finally, we sum them as the output of the integration decoder.

Fig. 4. The red rectangle in (a) is a fixed window which is unsuitable for some heads. The blue rectangle in (b) is the adaptive window which could well fit different head sizes.

3.5. Multi-scale SSIM loss

We first change the ground-truth to the FIDT map (Liang et al., 2022c) which could model the head relations using the Euclidean distance between head points and pixel points. It is expressed as:

$$F(x, y) = \frac{1}{D(x, y)^{(\gamma \times D(x,y) + \varphi)} + \xi} \tag{8}$$

where $D(x, y)$ denotes Euclidean distance between the pixel and the annotation point of its nearest head, γ, φ and ξ are the adjustment coefficients. Here, we set γ, φ, and ξ to 0.02, 0.75 and 1 respectively as Liang et al. (2022c).

SSIM (Wang et al., 2004) is a significant index to quantify the similarity between two images. And, for crowd localization (Liang et al., 2022c; Cao et al., 2018), it is treated as the loss function:

$$L_{SSIM}(E, G) = 1 - \frac{\left(2\mu_E\mu_G + \phi_1\right)\left(2\sigma_{EG} + \phi_2\right)}{\left(\mu_E^2 + \mu_G^2 + \phi_1\right)\left(\sigma_E^2 + \sigma_G^2 + \phi_2\right)} \tag{9}$$

where the predicted FIDT map and the ground-truth FIDT map are denoted as E and G respectively. The mean and variance are represented by μ and σ respectively, while ϕ_1 and ϕ_2 are constants.

To compute the SSIM loss, the existing methods for crowd localization (Liang et al., 2022c; Liu et al., 2019a) directly conduct on the whole image or apply a fixed window for each head position. However, the fixed-size window is unsuitable for various head sizes as shown in Fig. 4(a). Hence, we propose an adaptive window for each head to compute the SSIM loss, where the size of adaptive window depends on the neighbor distance between the heads as shown in Fig. 4(b). Furthermore, we select the adaptive windows at different scales to

fuse the multi-scale information. Hence, the formulated MsSSIM loss is proposed as:

$$L_{MsSSIM}(E,G) = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M} L_{SSIM}\left(E_{nm}, G_{nm}\right) \tag{10}$$

where N is the head counts, M is the number of scales, E_{nm} and G_{nm} are the regions of the nth head at the mth scale in the predicted and the ground-truth FIDT map, respectively. Note that the adaptive window of the mth scale is a square area whose side length is the m-nearest neighbor Euclidean distance of the nth head. For example, for a given head, the side length of the adaptive window at the 2nd scale is set to the average Euclidean distance of the two nearest heads of this given head. We calculate the average SSIM loss of the adaptive window at multiple scales as the loss value of this head, and the final MsSSIM loss value is obtained by averaging all head loss values. The proposed MsSSIM loss not only considers the various head size via the adaptive window, but also fuse the multi-scale information when computing the loss.

In summary, the total loss is expressed as:

$$L = L_{MSE}(E,G) + \eta L_{MsSSIM}(E,G) \tag{11}$$

where η is adjustment coefficient. L_{MSE} and L_{MsSSIM} could be treated as the global and local losses, respectively.

In the inference stage, we first obtain all the local maximum points in the predicted FIDT map through the 3×3 max-pooling, and then we set two adaptive thresholds T_{max} and T_{min}. When the local maximum is larger than T_{max}, this point is treated as a head. If the maximum value of FIDT map is less than T_{min}, we consider that there is no head in the whole image.

4. Experimental results

4.1. Datasets

We evaluate our method on five challenging public datasets.

ShanghaiTech (Zhang et al., 2016) consists of two parts: Part A and Part B. Part A contains 244,176 annotated heads, which has a count range of 33 to 3,139. The training set consists of 300 images and the test set consists of 182 images. Part B is composed of 88,488 annotated heads where the training set includes 400 images and the test set includes 316 images. The number of heads in this dataset ranges from 9 to 578.

UCF-QNRF (Idrees et al., 2018) includes 1,535 images with 1.25 million annotated heads. There are 1,201 images in the training set and 334 images in the test set. This dataset consists of a variety of scenes, whose count range is from 49 to 12,865.

JHU-Crowd++ (Sindagi et al., 2022) is composed of 4,372 images and 1.51 million annotations. The training, validation and test sets contain 2,272, 500 and 1,600 images, respectively. It includes the scenarios in severe weather and various lighting conditions, making it challenging. The total number of people in each image ranges from 0 to 25,791.

UCF_CC_50 (Idrees et al., 2013) contains 50 gray images with a total of 63,974 annotated heads. The number of heads for each image ranges from 94 to 4,543. This dataset contains dense crowd images. We utilize 5-fold cross-validation to evaluate the average test performance (Song et al., 2021; Liang et al., 2022c; Idrees et al., 2013).

NWPU-Crowd (Wang et al., 2021a) has a total of 5,109 images, including 3,109 training images, 500 validation images, and 1,500 test images. This dataset has 2.13 million annotated head points, and the number of people ranges from 0 to 20,033. It contains crowd images in a variety of scenarios including exposure, extreme darkness, and high density. The results on the test set come from the online evaluation benchmark website https://www.crowdbenchmark.com (Wang et al., 2021a).

4.2. Implementation details

VGG-16 (Simonyan and Zisserman, 2015) is the backbone, and the number of CSWin Transformer Block for each scale is set to 1. We set the scale number M in Eq. (10) to 3, and set the parameter η in Eq. (11) to 0.1. In the inference stage, T_{max} is λ_m times the maximum value of the FIDT map where λ_m is equal to 110/255, and T_{min} is set to 0.1. We keep the resolution for ShanghaiTech and UCF_CC_50. And for remaining datasets, we ensure that the image resolution does not exceed 2048×2048 while maintaining the original aspect ratio of the image. For ShanghaiTech, the batch size is set to 16, while for the remaining datasets, it is set to 8. Adam (Kingma and Ba, 2015) is utilized to optimize the model, setting the weight decay to $5e-4$ and the learning rate to $1e-4$.

4.3. Evaluation metrics

Following Idrees et al. (2018), Wang et al. (2021a), for the evaluation metrics of localization performance, we adopt Precision, Recall, and F1-measure. The predicted head point that falls within the distance threshold ψ from the ground-truth point is considered a True Positive, while those exceeding ψ are classified as False Positives. The ground-truth points which are not predicted are treated as False Negative. For ShanghaiTech, JHU-Crowd++ and UCF_CC_50, as in Liang et al. (2022b,c), we take $\psi_s = 4$ and $\psi_l = 8$ as two fixed thresholds, where $\psi_s = 4$ is the stricter one. For UCF-QNRF, we apply Precision, Recall and F1-measure across a range of thresholds from 1 to 100 as Abousamra et al. (2021), Liang et al. (2022c), Idrees et al. (2018). For NWPU-Crowd, following Liang et al. (2022b), Wang et al. (2021a), Wan et al. (2021), Lin and Chan (2023), ψ is determined by the real size of each head: $\psi = \sqrt{h^2 + w^2}/2$, where h represents the height and w represents the width of the head. For the evaluation of counting performance, we apply MAE and MSE following Liang et al. (2022b,c), Idrees et al. (2018), Wang et al. (2021a).

4.4. Comparison and analysis

Localization. As showcased in Tables 1–5, we conduct a comparison between our method and the state-of-the-art methods on five datasets to evaluate the localization performance.

For ShanghaiTech, UCF-QNRF, JHU-Crowd++ and UCF_CC_50, our method achieves the best localization performance. For the sparse dataset ShanghaiTech, compared to the state-of-the-art FIDTM (Liang et al., 2022c), our method achieves a higher F1-measure by 3.8% for ψ_s (a stricter setting) on Part A, and 8.3% improvement in F1-measure for ψ_s on Part B. For UCF-QNRF, a dense dataset, our method surpasses the state-of-the-art method OT-M (Lin and Chan, 2023) by 5.1% for Average F1-measure and reports the highest performance on all three metrics. For JHU-Crowd++, our method improves FIDTM (Liang et al., 2022c) by 2.0% F1-measure for ψ_s. For UCF_CC_50, our method outperforms FIDTM (Liang et al., 2022c) by 4.7% F1-measure for ψ_s. For NWPU-Crowd, our method demonstrates the highest Precision performance and achieves comparable results in terms of Recall and F1-measure. It should be noticed that our method only utilizes point-level annotations, whereas GMS (Wang et al., 2023a) uses box-level annotations. The results presented here prove the superior performance of our method for crowd localization.

As listed in Tables 1–5, our method achieves the highest performance on almost all datasets. We select two representative methods for analysis and comparison. Compared with Transformer-based method CLTR (Liang et al., 2022b) which only learns single scale information in the training process, our method considers multi-scale information fusion in both encoding and decoding stages. Compared with map-based method, i.e., FIDTM (Liang et al., 2022c), our method builds the long-range context dependencies on the combined feature maps and considers the multi-scale information fusion in the loss. Completed

Table 1
The performance (%) of localization on ShanghaiTech.

Methods	SHTech Part A						SHTech Part B					
	ψ_s			ψ_l			ψ_s			ψ_l		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
LCFCN (Laradji et al., 2018)	43.3	26.0	32.5	75.1	45.1	56.3	–	–	–	–	–	–
TopoCount (Abousamra et al., 2021)	41.7	40.6	41.1	74.6	72.7	73.6	63.4	63.1	63.2	82.3	81.8	82.0
LSC-CNN (Sam et al., 2021)	33.4	31.9	32.6	63.9	61.0	62.4	29.7	29.2	29.5	57.5	56.7	57.0
AutoScale (Xu et al., 2022)	56.2	54.2	55.2	74.4	71.7	73.0	–	–	–	–	–	–
CLTR (Liang et al., 2022b)	43.6	42.7	43.2	74.9	73.5	74.2	–	–	–	–	–	–
FIDTM (Liang et al., 2022c)	59.1	58.2	58.6	78.2	77.0	77.6	64.9	64.5	64.7	83.9	83.2	83.5
CSViT (Ours)	**63.5**	**61.3**	**62.4**	**80.3**	**77.6**	**78.9**	**73.4**	**72.6**	**73.0**	**87.2**	**86.3**	**86.7**

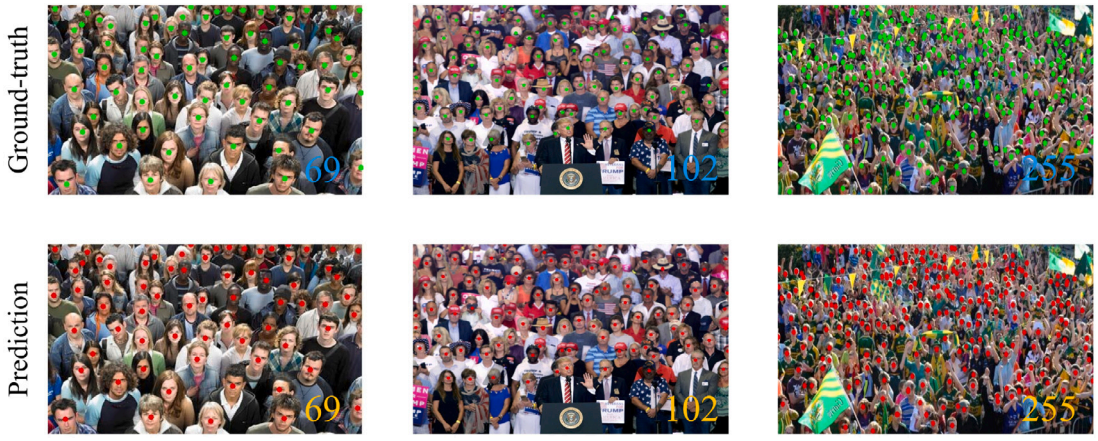

Fig. 5. Here are some qualitative results showcasing our crowd localization method. The ground-truth counts are indicated by blue numbers, while the predicted counts are represented by orange numbers.

Table 2
The performance (%) of localization on UCF-QNRF.

Methods	Av.Pre	Av.Rec	Av.F1
CL (Idrees et al., 2018)	75.8	59.8	66.8
LCFCN (Laradji et al., 2018)	77.9	52.4	62.7
LSC-CNN (Sam et al., 2021)	75.8	74.7	75.3
GL (Wan et al., 2021)	78.2	74.8	76.3
TopoCount (Abousamra et al., 2021)	81.8	79.0	80.3
AutoScale (Xu et al., 2022)	81.3	75.8	78.4
CLTR (Liang et al., 2022b)	82.2	79.8	81.0
FIDTM (Liang et al., 2022c)	84.5	80.1	82.2
OT-M (Lin and Chan, 2023)	80.4	78.3	79.3
CSViT (Ours)	**87.7**	**81.4**	**84.4**

Table 3
The performance (%) of localization on JHU-Crowd++.

Methods	ψ_s			ψ_l		
	Pre	Rec	F1	Pre	Rec	F1
TopoCount (Abousamra et al., 2021)	31.5	28.8	30.1	63.6	58.3	60.8
FIDTM (Liang et al., 2022c)	38.9	38.7	38.8	62.5	62.4	62.4
CSViT (Ours)	**41.5**	**40.2**	**40.8**	**64.6**	**62.6**	**63.6**

Table 4
The performance (%) of localization on UCF_CC_50.

Methods	ψ_s			ψ_l		
	Pre	Rec	F1	Pre	Rec	F1
LSC-CNN (Sam et al., 2021)	37.7	39.5	38.6	57.8	61.1	59.4
TopoCount (Abousamra et al., 2021)	39.5	42.0	40.7	62.5	66.9	64.6
AutoScale (Xu et al., 2022)	37.8	40.5	39.1	59.0	62.3	60.6
FIDTM (Liang et al., 2022c)	46.5	49.0	47.7	67.0	70.6	68.7
CSViT (Ours)	**52.4**	**52.6**	**52.4**	**70.6**	**70.9**	**70.7**

multi-scale information fusion and the long-range context dependencies enable our method to exceed these state-of-the-art methods.

Table 5
The performance (%) of localization on NWPU-Crowd. Bold and underline represent the best and second-place performance, respectively.

Methods	Pre	Rec	F1
TinyFaces (Hu and Ramanan, 2017)	52.9	61.1	56.7
RAZ_Loc (Liu et al., 2019d)	66.6	54.3	59.8
Crowd-SDNet (Wang et al., 2021c)	65.1	62.4	63.7
TopoCount (Abousamra et al., 2021)	69.5	68.7	69.1
SCALNet (Wang et al., 2021b)	69.2	69.0	69.1
GL (Wan et al., 2021)	<u>80.0</u>	56.2	66.0
AutoScale (Xu et al., 2022)	67.3	57.4	62.0
CLTR (Liang et al., 2022b)	69.4	67.6	68.5
FIDTM (Liang et al., 2022c)	79.7	<u>71.7</u>	75.5
OT-M (Lin and Chan, 2023)	71.0	65.8	68.3
GMS (Lin and Chan, 2023)	79.8	**76.5**	**78.1**
CSViT (Ours)	**82.9**	70.4	<u>76.1</u>

Furthermore, we give some visualization results of our method in Fig. 5. From the first and the second columns of this figure, we can see that the heads with large-scale variations are effectively localized. Additionally, our method achieves accurate localization results in the dense scene (column 3). These indicate that our method performs effective crowd localization in different scenarios.

Counting. Our method mainly focuses on crowd localization, and it could also obtain the crowd counting from the summation of localized heads. The comparison results with other crowd counting methods are displayed in the upper part of Table 6, where the blue bold indicates the best performance among crowd counting methods. Note that these compared methods only perform the crowd counting, while our method not only conducts the crowd counting, but also the crowd localization. By observing the table, it becomes evident that our method surpasses these crowd counting methods on all datasets except for MSE of ShanghaiTech Part A. This showcases the versatility of our method in handling various tasks, i.e., crowd counting and crowd localization.

Table 6
The performance of counting on ShanghaiTech, UCF-QNRF, JHU-Crowd++, UCF_CC_50 and NWPU-Crowd. Blue bold represents the best performance among crowd counting methods. Red bold and underline represent the best and second-place performance among crowd localization methods, respectively.

Methods	Position	SHTech Part A		SHTech Part B		QNRF		JHU		UCF_CC_50		NWPU	
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CSRNet (Li et al., 2018)	×	68.2	115.0	10.6	16.0	–	–	85.9	309.2	266.1	397.5	121.3	**387.8**
SFCN (Wang et al., 2019)	×	64.8	107.5	7.6	13.0	102.0	171.4	77.5	297.6	214.2	318.2	105.7	424.1
L2SM (Xu et al., 2019)	×	64.2	98.4	7.2	11.1	104.7	173.6	–	–	188.4	315.3	–	–
CG-DRCN (Sindagi et al., 2019)	×	64.0	98.4	8.5	14.4	112.2	176.3	82.3	328.0	–	–	–	–
DSSI-Net (Liu et al., 2019b)	×	60.6	96.0	6.9	10.3	99.1	159.2	133.5	416.5	216.9	302.4	–	–
MBTTBF (Sindagi and Patel, 2019)	×	60.2	94.1	8.0	15.5	97.5	165.2	81.8	299.1	233.1	300.9	–	–
BL (Ma et al., 2019)	×	62.8	101.8	7.7	12.7	88.7	154.8	75.0	299.9	229.3	308.2	105.4	454.2
RPNet (Yang et al., 2020)	×	61.2	96.9	8.1	11.6	–	–	–	–	–	–	–	–
ASNet (Jiang et al., 2020)	×	57.8	**90.1**	–	–	91.6	159.7	–	–	**174.8**	**251.6**	–	–
AMSNet (Hu et al., 2020)	×	56.7	93.4	**6.7**	**10.2**	101.8	163.2	–	–	208.6	296.3	–	–
LibraNet (Liu et al., 2020)	×	**55.9**	97.1	7.3	11.3	88.1	**143.7**	–	–	181.2	262.2	–	–
NoisyCC (Wan and Chan, 2020)	×	61.9	99.6	7.4	11.3	85.8	150.6	**67.7**	**258.5**	–	–	96.9	534.2
DM-Count (Wang et al., 2020)	×	59.7	95.7	7.4	11.8	**85.6**	148.3	–	–	211.0	291.5	**88.4**	388.6
DENet (Liu et al., 2021a)	×	65.5	101.2	9.6	15.4	121.0	205.0	–	–	241.9	345.4	–	–
DensityCNN (Jiang et al., 2021)	×	63.1	106.3	9.1	16.3	101.5	186.9	–	–	244.6	341.8	–	–
Lw-Count (Liu et al., 2022)	×	69.7	100.5	10.1	12.4	149.7	238.4	90.2	311.8	239.3	307.6	–	–
KDMG (Wan et al., 2022)	×	63.8	99.2	7.8	12.7	99.5	173.0	69.7	268.3	–	–	100.5	415.5
ECCNAS (Wang et al., 2022)	×	62.0	110.9	7.5	12.9	91.2	158.9	–	–	223.1	293.8	–	–
MPNet (Zhao et al., 2023)	×	65.4	108.4	9.3	13.4	–	–	–	–	–	–	–	–
CP-Net (Lyu et al., 2023)	×	58.5	95.4	6.7	10.6	91.2	156.6	–	–	198.2	283.9	–	–
PSDDN (Liu et al., 2019c)	✓	65.9	112.3	9.1	14.2	–	–	–	–	359.4	514.8	–	–
LSC-CNN (Sam et al., 2021)	✓	66.4	117.0	8.1	12.7	120.5	218.2	112.7	454.4	225.6	302.7	–	–
Crowd-SDNet (Wang et al., 2021c)	✓	65.1	104.4	7.8	12.6	–	–	–	–	–	–	–	–
TopoCount (Abousamra et al., 2021)	✓	61.2	104.6	7.8	13.7	89.0	159.0	60.9	267.4	184.1	258.3	107.8	438.5
GL (Wan et al., 2021)	✓	61.3	95.4	7.3	11.7	84.3	147.5	59.9	259.5	–	–	79.3	346.1
AutoScale (Xu et al., 2022)	✓	65.8	112.1	8.6	13.9	104.4	174.2	85.6	356.1	210.5	287.4	123.9	515.5
FIDTM (Liang et al., 2022c)	✓	57.0	103.4	6.9	11.8	89.0	153.5	66.6	253.6	171.4	233.1	86.0	**312.5**
CLTR (Liang et al., 2022b)	✓	56.9	95.2	6.5	10.6	85.8	141.3	–	–	–	–	**74.3**	333.8
GMS (Wang et al., 2023a)	✓	68.8	138.6	16.0	33.5	104.4	197.4	70.2	316.8	–	–	84.7	361.5
CSViT (Ours)	✓	**51.5**	**92.9**	**6.0**	**10.0**	**79.5**	**141.2**	**58.0**	**198.8**	**161.3**	**212.2**	75.8	330.5

Table 7
The effectiveness of combining multi-scale feature maps of CsViT on ShanghaiTech Part A, where the measurement unit of Pre, Rec and F1 is %.

Methods	MAE	MSE	ψ_s			ψ_l		
			Pre	Rec	F1	Pre	Rec	F1
BL	57.6	103.2	59.1	58.5	58.9	77.9	76.6	77.2
BL+IE	54.3	96.3	62.1	60.1	61.6	78.9	77.0	78.0
BL+ID	54.5	97.5	61.8	59.7	61.2	78.6	76.8	77.8
BL+IE+ID	**51.5**	**92.9**	**63.5**	**61.3**	**62.4**	**80.3**	**77.6**	**78.9**

Table 8
The effectiveness of the proposed MsSSIM loss on ShanghaiTech Part A, where the measurement unit of Pre, Rec and F1 is %.

Methods	MAE	MSE	ψ_s			ψ_l		
			Pre	Rec	F1	Pre	Rec	F1
MSE	59.6	101.5	60.5	58.9	60.0	77.8	76.6	77.2
MSE+SSIM	56.4	97.5	61.7	59.5	61.2	78.3	76.9	77.9
MSE+SSIM*	54.1	96.3	62.3	60.3	61.6	78.9	77.1	78.2
MSE+MsSSIM	**51.5**	**92.9**	**63.5**	**61.3**	**62.4**	**80.3**	**77.6**	**78.9**

We compare the counting performance with the state-of-the-art crowd localization methods as depicted in the bottom part of Table 6. Our method achieves the highest level of performance on all datasets except for NWPU-Crowd. It is essential to mention that our method significantly boosts counting accuracy on all datasets compared to GMS (Wang et al., 2023a). In particular, on ShanghaiTech Part B, the improvements are 62.5% and 70.1% in MAE and MSE, respectively. The findings indicate that our method exhibits a more comprehensive and well-balanced performance in both counting and localization.

4.5. Ablation studies

To verify the effectiveness of key components of our method, we perform the ablation studies on ShanghaiTech Part A.

Effectiveness of combining multi-scale feature maps. Table 7 shows the comparison results where BL is the baseline, and IE and ID represent the integration encoder and the integration decoder, respectively. Note that BL is obtained by removing the integration encoder and the integration decoder from the proposed CsViT. From the table, we can see that the performance of BL+IE and BL+ID is better than that of BL because multi-scale information fusion is considered in the encoding or decoding stage. Furthermore, the performance of BL+IE+ID is the best among all compared methods because it considers completed multi-scale information fusion. Specifically, for the counting performance, MAE and MSE of BL+IE+ID are lower than those of BL by a large margin of 6.1 and 10.3, respectively. For the localization performance, BL+IE+ID improves BL by 3.5% F1-measure (ψ_s) and 1.7% F1-measure (ψ_l), respectively.

Effectiveness of MsSSIM loss. Our investigation focuses on evaluating the contribution of the proposed MsSSIM loss. The corresponding results are shown in Table 8, where SSIM directly computes the loss on the whole image and SSIM* computes the loss for each head using the fixed-size head window (30×30) (Liang et al., 2022c). Firstly, the performance is improved when adding the SSIM loss on the MSE loss. The observed phenomenon could be attributed to the application of SSIM loss, which effectively quantifies the deviation between the predicted FIDT map and the actual FIDT map. Secondly, the performance of MSE+SSIM* is better than that of MSE+SSIM because SSIM* could partially mitigate the adverse impact caused by the background. Thirdly, the performance of MSE+MsSSIM is the best, which demonstrates that adaptive windows at different scales could precisely calculate the loss values of head regions for crowd localization and crowd counting.

4.6. Parameter analysis

We analyze the effect of the number of scales in CsViT and the number of scales M in MsSSIM. We list the effect of the scale number of the proposed CsViT in Table 9, showcasing the enhancement in

Table 9
The effect of the number of scales in CsViT on ShanghaiTech Part A, where the measurement unit of Pre, Rec and F1 is %.

Scales	MAE	MSE	ψ_s			ψ_l		
			Pre	Rec	F1	Pre	Rec	F1
2	57.1	101.0	60.7	58.7	60.3	77.9	76.7	77.4
3	54.4	96.5	61.9	60.6	61.3	78.5	77.3	77.9
4	**51.5**	**92.9**	**63.5**	**61.3**	**62.4**	**80.3**	**77.6**	**78.9**

Table 10
The effect of the number of scales M in MsSSIM on ShanghaiTech Part A, where the measurement unit of Pre, Rec and F1 is %.

M	MAE	MSE	ψ_s			ψ_l		
			Pre	Rec	F1	Pre	Rec	F1
2	53.5	95.8	62.8	60.6	61.9	79.2	77.3	78.5
3	**51.5**	**92.9**	**63.5**	**61.3**	**62.4**	**80.3**	**77.6**	**78.9**
4	53.9	95.5	63.0	60.4	62.2	79.5	77.1	78.6
5	54.0	96.7	62.5	60.2	62.0	78.9	76.8	78.4

counting and localization performance with an increasing number of scales. It indicates that the combination of feature maps across multiple scales could effectively deal with multi-scale heads. Hence, we choose to combine the feature maps of all four scales of VGG-16.

Table 10 shows the effect of the number of scales M in MsSSIM. Too few scales could not robustly represent the adaptive window of the head, and too many scales introduce outliers, i.e., large or minimal the head neighbor distance. Therefore, as shown in Table 10, a suitable number of scales, i.e., $M = 3$, achieves the best performance.

4.7. Limitations

The potential drawback of the proposed method is its computational complexity. The computational complexity of our method is 578.7 GMACs and the inference speed is 7.1 FPS. The experiments are conducted on a 3090 GPU, and the size of the input image is 1024×768. In practical applications, the crowd localization requires faster inference speed. Thus, our subsequent direction is to develop a lightweight deep model while maintaining the localization and counting performance in order to achieve real-time in the inference stage.

5. Conclusion

In this paper, we have proposed CsViT for crowd localization, which constructs completed multi-scale information fusion strategy in the encoding and decoding stages. Furthermore, we propose the MsSSIM loss to calculate the SSIM loss for each head using the adaptive windows at different scales. We have performed extensive experiments on five publicly available datasets, and the experimental results prove that our method achieves the state-of-the-art performance in the crowd localization and counting. In the future, we will develop the deep model with low sensitivity to hyperparameters and focus on the interpretability of the model.

CRediT authorship contribution statement

Shuang Liu: Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Project administration. **Yu Lian:** Software, Formal analysis, Writing – original draft. **Zhong Zhang:** Formal analysis, Writing – review & editing, Supervision. **Baihua Xiao:** Validation, Formal analysis, Writing – review & editing. **Tariq S. Durrani:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by National Natural Science Foundation of China under Grant No. 62171321, Natural Science Foundation of Tianjin, China under Grant No. 22JCQNJC00010, Scientific Research Project of Tianjin Educational Committee, China under Grant No. 2022KJ011, and Tianjin Normal University Research Innovation Project for Postgraduate Students, China under Grant No. 2023KYCX003Z.

References

Abousamra, S., Hoai, M., Samaras, D., Chen, C., 2021. Localization in the crowd with topological constraints. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 872–881.

Basalamah, S., Khan, S.D., Felemban, E., Naseer, A., Rehman, F.U., 2023. Deep learning framework for congestion detection at public places via learning from synthetic data. J. King Saud Univ. - Comput. Inf. Sci. 35 (1), 102–114.

Cao, X., Wang, Z., Zhao, Y., Su, F., 2018. Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European Conference on Computer Vision. pp. 734–750.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision. pp. 213–229.

Chen, Z., Joseph Raj, A.N., Rajangam, V., Li, W., Mahesh, V.G., Zhuang, Z., 2023. Twofold dynamic attention guided deep network and noise-aware mechanism for image denoising. J. King Saud Univ. - Comput. Inf. Sci. 35 (3), 87–102.

Deng, M., Zhao, H., Gao, M., 2023. CLFormer: a unified transformer-based framework for weakly supervised crowd counting and localization. Vis. Comput. 1–15.

Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B., 2022. CSWin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2021. An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations. pp. 1–21.

Gong, Y., Zhang, Y., Cun, X., Yin, F., Fan, Y., Wang, X., Wu, B., Yang, Y., 2023. ToonTalker: Cross-domain face reenactment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7690–7700.

He, F., Huang, Y., Wang, H., 2022. IPLAN: Interactive and procedural layout planning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7793–7802.

Hu, Y., Jiang, X., Liu, X., Zhang, B., Han, J., Cao, X., Doermann, D., 2020. NAS-count: Counting-by-density with neural architecture search. In: Proceedings of the European Conference on Computer Vision. pp. 747–766.

Hu, P., Ramanan, D., 2017. Finding tiny faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 951–959.

Idrees, H., Saleemi, I., Seibert, C., Shah, M., 2013. Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2547–2554.

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M., 2018. Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European Conference on Computer Vision. pp. 532–546.

Jiang, X., Zhang, L., Xu, M., Zhang, T., Lv, P., Zhou, B., Yang, X., Pang, Y., 2020. Attention scaling for crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4706–4715.

Jiang, X., Zhang, L., Zhang, T., Lv, P., Zhou, B., Pang, Y., Xu, M., Xu, C., 2021. Density-aware multi-task learning for crowd counting. IEEE Trans. Multimed. 23, 443–453.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations. pp. 1–15.

Laradji, I.H., Rostamzadeh, N., Pinheiro, P.O., Vazquez, D., Schmidt, M., 2018. Where are the blobs: Counting by localization with point supervision. In: Proceedings of the European Conference on Computer Vision. pp. 547–562.

Li, Y., Zhang, X., Chen, D., 2018. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1091–1100.

Lian, D., Li, J., Zheng, J., Luo, W., Gao, S., 2019. Density map regression guided detection network for RGB-D crowd counting and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1821–1830.

Liang, D., Chen, X., Xu, W., Zhou, Y., Bai, X., 2022a. Transcrowd: weakly-supervised crowd counting with transformers. Sci. China Inf. Sci. 65 (6), 160104.

Liang, D., Xie, J., Zou, Z., Ye, X., Xu, W., Bai, X., 2023. CrowdCLIP: Unsupervised crowd counting via vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2893–2903.

Liang, D., Xu, W., Bai, X., 2022b. An end-to-end transformer model for crowd localization. In: Proceedings of the European Conference on Computer Vision. pp. 38–54.
Liang, D., Xu, W., Zhu, Y., Zhou, Y., 2022c. Focal inverse distance transform maps for crowd localization. IEEE Trans. Multimed. 1–13.
Lin, W., Chan, A.B., 2023. Optimal transport minimization: Crowd localization on density maps for semi-supervised counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21663–21673.
Lin, Y., Huang, J., Sun, D., 2023. A novel recurrent convolutional network based on grid correlation modeling for crowd flow prediction. J. King Saud Univ. - Comput. Inf. Sci. 35 (8), 101699.
Liu, Y., Cao, G., Shi, H., Hu, Y., 2022. Lw-count: An effective lightweight encoding-decoding crowd counting network. IEEE Trans. Circuits Syst. Video Technol. 32 (10), 6821–6834.
Liu, J., Gao, C., Meng, D., Hauptmann, A.G., 2018. DecideNet: Counting varying density crowds through attention guided detection and density estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5197–5206.
Liu, L., Jiang, J., Jia, W., Amirgholipour, S., Wang, Y., Zeibots, M., He, X., 2021a. DENet: A universal network for counting crowd with varying densities and scales. IEEE Trans. Multimed. 23, 1060–1068.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.
Liu, L., Lu, H., Zou, H., Xiong, H., Cao, Z., Shen, C., 2020. Weighing counts: Sequential crowd counting by reinforcement learning. In: Proceedings of the European Conference on Computer Vision. pp. 164–181.
Liu, S., Peng, W., Liu, Y., Zhao, J., Su, Y., Zhang, Y., 2023. AFCANet: An adaptive feature concatenate attention network for multi-focus image fusion. J. King Saud Univ. - Comput. Inf. Sci. 35 (9), 101751.
Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., Lin, L., 2019a. Crowd counting with deep structured scale integration network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1774–1783.
Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., Lin, L., 2019b. Crowd counting with deep structured scale integration network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1774–1783.
Liu, Y., Shi, M., Zhao, Q., Wang, X., 2019c. Point in, box out: Beyond counting persons in crowds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6469–6478.
Liu, C., Weng, X., Mu, Y., 2019d. Recurrent attentive zooming for joint crowd counting and precise localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1217–1226.
Lyu, L., Han, R., Chen, Z., 2023. Cascaded parallel crowd counting network with multi-resolution collaborative representation. Appl. Intell. 53 (3), 3002–3016.
Ma, Z., Wei, X., Hong, X., Gong, Y., 2019. Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6142–6151.
Qiu, J., Wan, W., Yao, H., Han, K., 2017. Crowd counting and density estimation via two-column convolutional neural network. In: Proceedings of the International Conference on Smart and Sustainable City. pp. 1–5.
Sam, D.B., Peri, S.V., Sundararaman, M.N., Kamath, A., Babu, R.V., 2021. Locate, size and count: Accurately resolving people in dense crowds via detection. IEEE Trans. Pattern Anal. Mach. Intell. 43 (8), 2739–2751.
Si, T., He, F., Li, P., Song, Y., Fan, L., 2023. Diversity feature constraint based on heterogeneous data for unsupervised person re-identification. Inf. Process. Manage. (ISSN: 0306-4573) 60 (3), 103304.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations. pp. 1–14.
Sindagi, V.A., Patel, V.M., 2019. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1002–1012.
Sindagi, V.A., Yasarla, R., Patel, V.M., 2019. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1221–1231.
Sindagi, V.A., Yasarla, R., Patel, V.M., 2022. JHU-CROWD++: Large-scale crowd counting dataset and a benchmark method. IEEE Trans. Pattern Anal. Mach. Intell. 44 (5), 2594–2609.
Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Wu, Y., 2021. Rethinking counting and localization in crowds: A purely point-based framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3365–3374.
Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021. Training data-efficient image transformers & distillation through attention. In: Proceedings of the International Conference on Machine Learning. pp. 10347–10357.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems. pp. 5998–6008.
Wan, J., Chan, A.B., 2020. Modeling noisy annotations for crowd counting. In: Proceedings of Advances in Neural Information Processing Systems. pp. 3386–3396.
Wan, J., Liu, Z., Chan, A.B., 2021. A generalized loss function for crowd counting and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1974–1983.
Wan, J., Wang, Q., Chan, A.B., 2022. Kernel-based density map generation for dense object counting. IEEE Trans. Pattern Anal. Mach. Intell. 44 (3), 1357–1370.
Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.
Wang, Q., Gao, J., Lin, W., Li, X., 2021a. NWPU-crowd: A large-scale benchmark for crowd counting and localization. IEEE Trans. Pattern Anal. Mach. Intell. 43 (6), 2141–2149.
Wang, Q., Gao, J., Lin, W., Yuan, Y., 2019. Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8198–8207.
Wang, J., Gao, J., Yuan, Y., Wang, Q., 2023a. Crowd localization from Gaussian mixture scoped knowledge and scoped teacher. IEEE Trans. Image Process. 32, 1802–1814.
Wang, R., Hao, Y., Hu, L., Chen, J., Chen, M., Wu, D., 2023b. Self-supervised learning with data-efficient supervised fine-tuning for crowd counting. IEEE Trans. Multimed. 25, 1538–1546.
Wang, Y., Hou, X., Chau, L.-P., 2021b. Dense point prediction: A simple baseline for crowd counting and localization. In: Proceedings of the IEEE International Conference on Multimedia Expo Workshops. pp. 1–6.
Wang, Y., Hou, J., Hou, X., Chau, L.-P., 2021c. A self-training approach for point-supervised object detection and counting in crowds. IEEE Trans. Image Process. 30, 2876–2887.
Wang, B., Liu, H., Samaras, D., Nguyen, M.H., 2020. Distribution matching for crowd counting. In: Proceedings of Advances in Neural Information Processing Systems. pp. 1595–1607.
Wang, Y., Ma, Z., Wei, X., Zheng, S., Wang, Y., Hong, X., 2022. Eccnas: Efficient crowd counting neural architecture search. ACM Trans. Multimed. Comput. Commun. Appl. 18 (1), 1–19.
Wang, X., Zhan, Y., Zhao, Y., Yang, T., Ruan, Q., 2023c. Semi-supervised crowd counting with spatial temporal consistency and pseudo-label filter. IEEE Trans. Circuits Syst. Video Technol. 33 (8), 4190–4203.
Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L., 2021. CvT: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22–31.
Xu, C., Liang, D., Xu, Y., Bai, S., Zhan, W., Bai, X., Tomizuka, M., 2022. Autoscale: learning to scale for crowd counting. Int. J. Comput. Vis. 130 (2), 405–434.
Xu, C., Qiu, K., Fu, J., Bai, S., Xu, Y., Bai, X., 2019. Learn to scale: Generating multipolar normalized density maps for crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8382–8390.
Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., Sebe, N., 2020. Reverse perspective network for perspective-aware object counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4374–4383.
Zhang, X., Zhang, S., Cui, Z., Li, Z., Xie, J., Yang, J., 2023. Tube-embedded transformer for pixel prediction. IEEE Trans. Multimed. 25, 2503–2514.
Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y., 2016. Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 589–597.
Zhao, Z., Li, X., 2023. Deformable density estimation via adaptive representation. IEEE Trans. Image Process. 32, 1134–1144.
Zhao, H., Wang, Q., Zhan, G., Min, W., Zou, Y., Cui, S., 2023. Need only one more point (NOOMP): Perspective adaptation crowd counting in complex scenes. IEEE Trans. Multimed. 25, 1414–1426.