

Depth Guided Scale-Aware Transformer for Crowd Counting

Qijun Lu^[0009-0003-9304-0166] and Jinhua Xu^{(✉) [0000-0002-8450-3144]}

School of Computer Science and Technology, East China Normal University,
Shanghai 200062, China
jhxu@cs.ecnu.edu.cn

Abstract. Crowd counting is an important task in computer vision with wide applications. Deep learning methods have achieved great progress in the crowd counting task, but scale variation in the crowd images is still a challenge. To address this issue, we propose a Depth Guided Scale-Aware Transformer (DGSAT) for crowd counting in this paper. Three novel modules are designed including Dilated Rectangle Window Attention (DRWA), Depth-guided Scale Segmentation (DGSS), and Size-Aware Hungarian Matching (SAHM). In the DRWA module, we apply different dilation rates to the conventional window attention mechanism to collect context from different scales. In the DGSS module, the depth map is used to guide the prediction of a scale segmentation mask so that features from different scales can be selected to localize the heads of different scales. In the SAHM module, head sizes are estimated and used to balance the position error and the classification confidence score in the cost matrix of the Hungarian matching. Our method has been extensively evaluated on crowd counting datasets and achieves SOTA results on the ShanghaiTech Part A dataset. The source code is available at <https://github.com/LuqijunLab/DGSAT>.

Keywords: Crowd Counting, Transformer, Window Attention, Hungarian Matching.

1 Introduction

Crowd counting is a critical task in computer vision, with applications in public safety, urban planning, event management, and epidemic control [1,2]. The number of people in an image or a video provides essential data for decision making in scenarios such as monitoring crowd density in public spaces or optimizing resource allocation during large-scale events. Recent advances in deep learning have achieved great success for crowd counting. The methods can be categorized into density map-based methods [3–6] and point-based methods [7–10].

The crowd counting task faces many challenges, such as occlusion, complex background, scale variation, uneven distribution, etc. Among them, the intra-image scale variations caused by the perspective-related distortions shown in Fig. 1 have been drawing extensive attention. Convolutional Neural Networks (CNNs) are widely used in



Fig. 1. Scale variation in crowd images.

crowd counting tasks due to their ability of multi-scale feature learning. Traditional methods for multi-scale feature generation include extracting features from multiple layers (levels) [11] of CNNs, multi-branches (columns) [12,13] with varied receptive field sizes. However, such schemes usually suffer from several limitations such as redundant parameters, huge computation burden and difficulty in optimization [14]. Dilated convolutions have been used to increase the receptive fields to better leverage multi-scale information while keeping the resolution and without extra computation cost [14–17]. In [17], a Scale Tree Network (STNet) is proposed to enrich scale diversity using various dilation rates and to achieve the hierarchical parsing of coarse-to-fine scales, which alleviates limitations caused by insufficient level of scales. In [16], an adaptive dilated convolution is proposed, in which a continuous value is predicted as the specific dilation rate for each location, which adapts the scale variation better than a discrete and static dilation rate. Similarly, Perspective-guided Fractional-Dilation Network (PFDNet) is proposed in [14] for modeling the continuous scale variations.

Transformers have also been applied to crowd counting for their strong capability in modeling long dependencies and global context. There is no scale limitation due to the global context modeling of the vanilla transformer. However, the global attention mechanism incurs high computational complexity. Window attention is introduced in Swin Transformer [18], which significantly reduces computational overhead. TransCrowd [19] incorporates an additional token in the Transformer encoder to directly predict the total crowd count without explicitly locating individual head points. Transformers have also been used in the point-based method for crowd counting [8,9]. The crowd counting problem is viewed as a point set prediction problem. However, we found an unstable matching issue caused by the scale variation. The cost function of Hungarian Match consists of two parts: the position error and the classification confidence score. As shown in Fig. 2, the annotated points in crowd images often exhibit positional offsets and are not precisely centered on the heads. For large scale heads, the localization error may be dominant and overwhelm the confidence score.

In this paper, we propose a point-based crowd counting method using the transformer framework. We tackle the scale variation problem from the following aspects. First, inspired by the dilated attention mechanism, a Dilated Rectangle Window Attention Mechanism is proposed to introduce progressively longer dependencies through varying dilation rates without extra computational overhead. Second, crowd images



Fig. 2. The annotated points often deviate from the head centers.

contain inherent depth information, which is a primary factor contributing to scale variation. To effectively leverage the depth information, a Depth-Guided Scale Segmentation (DGSS) module is designed. This module uses a depth estimation model to predict a depth map, which is used to generate mask labels. The DGSS module learns a scale mask to distinguish between near-range and far-range regions, then features from appropriate scales can be dynamically selected to improve localization precision. Third, to solve unstable matching problem, a Size-Aware Hungarian Matching strategy is designed. We estimate the head sizes and dynamically adjust the matching weights according to the head sizes. This strategy balances localization errors and the classification score, enhancing the robustness and accuracy of the matching process.

In summary, the main contributions of our work include: 1) Dilated Rectangle Window Attention (DRWA) Mechanism. By employing varying dilation rates, DRWA expands the attended receptive fields without increasing the computation cost. 2) Depth-Guided Scale Segmentation (DGSS). A scale mask is learned based on the depth to select features from appropriate scales. 3) Size-Aware Hungarian Matching (SAHM). A novel size-aware matching approach is introduced to dynamically adjust matching weights based on head sizes.

The remainder of this paper is organized as follows. Section 2 reviews related work on crowd counting tasks. Section 3 introduces the proposed method in detail. Section 4 presents experimental results on benchmark datasets and comparisons with the existing approaches. The conclusion is given in Section 5.

2 Related work

This section reviews related work on crowd counting tasks, including density map-based methods, point-based methods, window attention mechanism and depth-based counting methods.

2.1 Density Map-Based Methods

Density map-based methods convert the discrete crowd counting problem into a continuous density regression problem [3–6]. Given an image I with N annotated head points $\{P_1, P_2, \dots, P_N\}$, a Gaussian kernel is applied to each point P_i to generate the

Gaussian distribution. These distributions are summed to form the image’s density map label, and the predicted count is obtained by summing over the predicted density map. In [20,21], a depth-adaptive Gaussian kernel is introduced to generate high-fidelity density map for more robust density map regression. In [22], low-rank Gaussian convolution is adopted instead of traditional convolution to mitigate noise and spatial invariance in generated density maps. A Focal Inverse Distance Transform (FIDT) map is proposed in [23] which accurately describes the persons’ locations without overlapping in dense regions. Most of density map-based methods [12,24] are CNN-based and can be mainly divided into two categories, multi-scale features fusion and multi-scale prediction fusion. The former fuses multi-scale features first and then predicts one density map [11,24]. The latter generates one prediction for each scale feature and then combines the multiple predictions [5,13].

2.2 Point-Based Methods

Point-based methods directly use annotated points as the learning target to avoid inaccuracies in adapting to varying head scales during density map generation [7–10]. P2PNet [7] introduces a point-to-point prediction framework that uses CNN features to predict candidate points and their confidence scores, followed by Hungarian matching to identify optimal matches between point proposals and ground truth points. The cost matrix is composed of L_2 point distances and point confidence scores. Liang et al. [8] propose a crowd localization transformer (CLTR), which views the crowd localization as a direct set prediction problem, and the average distance between a ground truth point (or a proposal point) and its neighbors is incorporated in the cost matrix to improve point matching consistency. Liu et al. [9] propose a Point Query Transformer (PET) for crowd counting, featured by the point-query quadtree and progressive rectangle window attention. PET implements decomposable point querying via data-dependent quadtree splitting, where each querying point could split into four new points when necessary, thus enabling dynamic processing of sparse and dense regions. APGCC [10] alleviates unstable matching during training by sampling positive and negative auxiliary points around labeled points as additional learning targets.

In this study, we propose a point-based method using the transformer architecture. A Size-Aware Hungarian Matching (SAHM) module is designed to enhance the robustness and accuracy of the matching process.

2.3 Window Attention Mechanism

The global attention mechanism in standard Transformer models incurs high computational complexity. Swin Transformer [18] introduces shifted window attention, which computes attention within fixed-size local windows and shifts window positions across layers. This localized attention mechanism significantly reduces computational overhead. DiNAT [25] applies dilation to the attention mechanism to further expand the receptive field and capture long-range context without adding any additional computational burden. Dilateformer [26] introduces the Sliding Window Dilated Attention (SWDA), where different dilation rates are applied to the attention heads within each

block. In crowd counting tasks, PET [9] employs a progressive window attention mechanism, dividing feature maps into rectangular regions of varying sizes across layers.

In this work, we introduce a dilated rectangle window attention mechanism to introduce progressively longer dependencies through varying dilation rates without extra computational overhead.

2.4 Depth Information for Crowd Counting

Incorporating depth information into crowd counting has shown promising results in addressing challenges like occlusion, perspective distortion, and scale variation [20, 21, 27–30]. Lian et al. [20, 21] propose a depth-guided detection module to generate a dynamic dilated convolution to extract features of different scales, and a depth-aware anchor is further designed for better initialization of anchor sizes in the detection framework. DECCNet [30] introduces a bidirectional cross-modal attention mechanism to fuse RGB and depth features, enabling the model to prioritize congested regions based on depth cues. DigCrowd [29] uses the depth information to segment the scene into a far-view region and a near-view region, then applies a detection method to count the people in the near-view region and a density map method in the far-view region. These works demonstrate that depth information helps resolve ambiguities in crowded scenes by encoding 3D spatial relationships and scale variations.

Recent advancements in monocular depth estimation have significantly enhanced the ability to infer depth information from 2D images. DepthAnything [31] has emerged as a robust framework, combining multi-scale feature fusion and adaptive loss functions to achieve superior depth estimation across varying illumination and scene complexities. In this work, we present a depth-guided crowd counting framework that leverages the DepthAnything [31] model to generate high-quality depth maps. Unlike prior methods that use depth for density map generation [20,21], feature fusion [30] or method selection [29], we design a lightweight network to dynamically learn depth-guided scale masks for feature scale selection. In addition, we leverage depth information to assist in head size estimation, thereby mitigating the impact of location errors for large scale heads during matching process.

3 Method

In this section, we first introduce the architecture of our proposed method. Then, we will introduce three novel modules: Dilated Rectangle Window Attention, Depth-guided Scale Segmentation, and Size-Aware Hungarian Matching. Finally, we present the loss function.

3.1 Architecture Overview

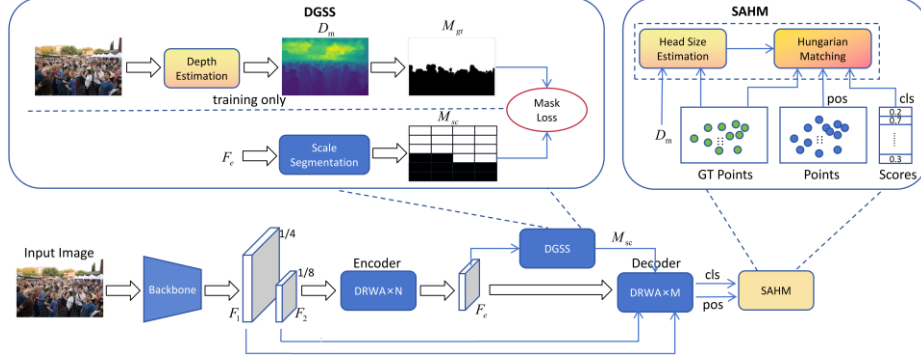


Fig. 3. Overview of the proposed model. It consists of backbone, encoder, decoder, Depth-Guided Scale Segmentation (DGSS) module and Size-aware Hungarian matching (SAHM) module. Dilated Rectangle Window Attention (DRWA) is used in the encoder and the decoder. The modules in blue contain parameters to be trained while the ones in yellow do not have learnable parameters and are used in the training stage only.

The architecture of the proposed method is illustrated in Fig. 3. The model consists of the backbone, encoder, decoder, depth guided scale segmentation (DGSS) module and scale aware Hungarian Matching (SAHM) module. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, the backbone is used to extract multi-scale features from I . The final two stage feature maps, $F_1 \in \mathbb{R}^{\frac{1}{4}H \times \frac{1}{4}W \times c}$ and $F_2 \in \mathbb{R}^{\frac{1}{8}H \times \frac{1}{8}W \times c}$, are selected to represent low-level and high-level features, respectively, where c denotes the feature dimension, H and W are the height and the width of the input image respectively.

The extracted features from the backbone network are further processed using a Transformer-based structure enhanced with our proposed Dilated Rectangle Window Attention (DRWA) module, as illustrated in Fig. 4. During the encoding stage, the high-level feature map F_2 is first partitioned into a series of rectangular windows based on a specified dilation rate d . These windows are then fed into a Transformer encoder layer. This process is repeated N times with different dilation rates, and the final output is the encoded feature F_e .

In the decoding stage, the low-level feature F_1 and high-level feature F_2 from the backbone are used as the query Q , while the encoded feature F_e serves as the key K and the value V . The features are first partitioned into a series of rectangular windows based on a specified dilation rate d , and then passed into a Transformer decoder layer. This process is repeated M times using different dilation rates. The output features of the decoder are fed into a feed-forward network (FFN) for head classification (cls) and head position (pos) regression. Specifically, a linear layer is used for head classification, while a MLP consisting of three linear layers with ReLU activations in between is used for head position regression.

Additionally, the encoded feature F_e is input into the DGSS module to predict a scale mask M_{sc} , which is used to select the corresponding features from F_1 and F_2 as the query Q of the decoder.

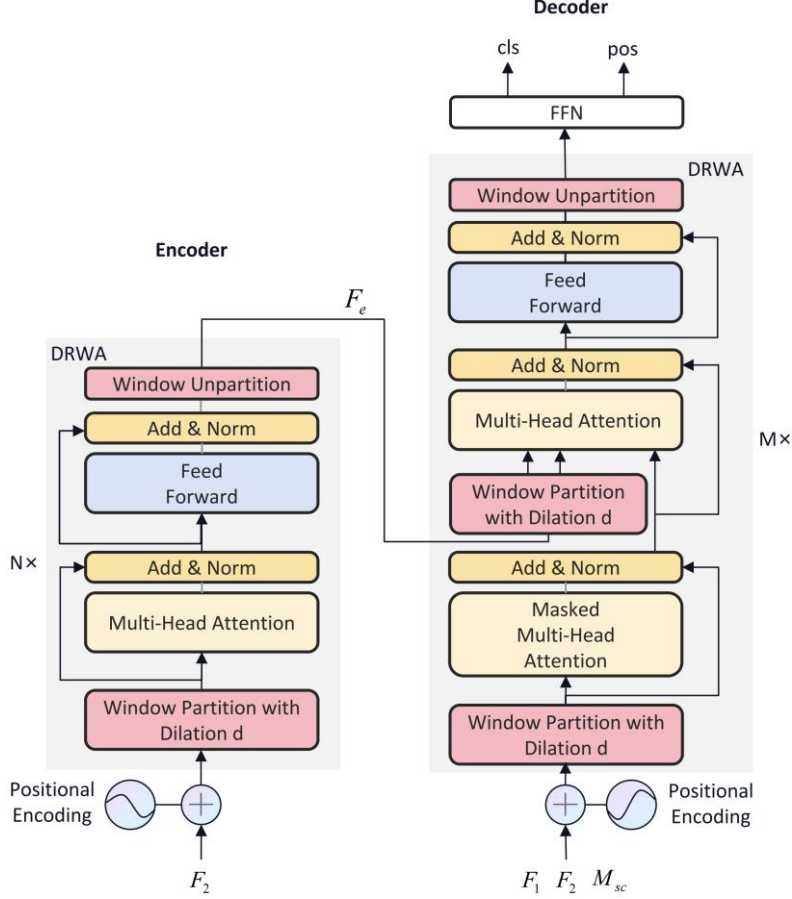


Fig. 4. The Transformer-based structure enhanced with our proposed Dilated Rectangle Window Attention (DRWA) module, where the features are partitioned into a series of rectangular windows based on a specified dilation rate d .

3.2 Dilated Rectangle Window Attention (DRWA)

Rectangle Window Attention mechanism is used in PET [9], which demonstrates promising results in crowd counting tasks. Window attention effectively reduces computational complexity. However, using small windows may result in the loss of context information beyond window boundaries. Inspired by the dilated attention mechanism [25,26], we propose a Dilated Rectangle Window Attention module, which extends the small rectangle windows to achieve larger receptive fields. In our Encoder and Decoder of the transformer, we use multiple DRWA modules with different dilation rates (e.g., 4, 2, 1) to capture contextual information from different scales. DRWA is essentially an extension of the vanilla rectangle window attention mechanism. When $d = 1$, it

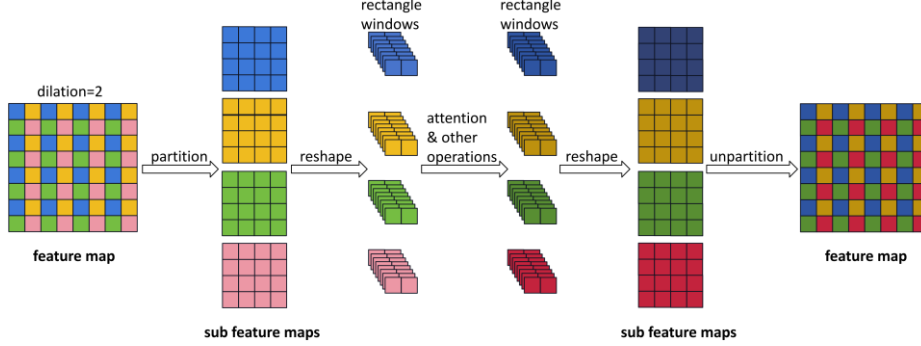


Fig. 5. The computation of Dilated Rectangle Window Attention (DRWA). Here the dilation rate is set to 2.

corresponds to the window attention mechanism. Given a feature map $F \in \mathbb{R}^{H \times W \times C}$ and a dilation rate d , we first select features along the height (H) and width (W) dimensions with a stride of d , resulting in $d \times d$ sub-feature maps. Each sub-feature map is then partitioned into a series of rectangular windows for attention computation. The results are merged to the sub-feature map shape and finally restored to their corresponding positions in the original feature map. Fig. 5 illustrates the computation process of DRWA.

3.3 Depth-Guided Scale Segmentation (DGSS)

A Crowd images contain depth information implicitly since the head size in the image changes with the depth. Therefore, we segment different scale regions according to the depth. Motivated by previous works [29,30], we employ a depth prediction model [31] to estimate a depth map $D \in \mathbb{R}^{H \times W}$ from the input image I . D is represented as a gray-scale image, with depth values normalized to the range (0,1). To separate near-range and far-range regions, we define a threshold d_{th} , where regions with the depth greater (less) than d_{th} are classified as the far(near)-range. We then obtain the ground truth scale mask M_{gt} as follows:

$$M_{gt} = \begin{cases} 1, & \text{if } AdaptAvg(D) > d_{th}, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

AdaptAvg refers to the adaptive average pooling operation that adjusts the size of the depth map to match the dimensions of the predicted mask M_{sc} . For far regions where head sizes are smaller, lower-level feature map F_1 is used for counting. For near regions where head sizes are larger, high-level feature map F_2 is used.

We use the output of the encoder as the input of a lightweight scale segmentation network to estimate the scale mask M_{sc} . The network consists of only an average pooling layer and a convolutional layer, hence it introduces minimal additional computational overhead. It should be noted that the depth prediction model is used only during

training to generate the ground truth scale mask M_{gt} to supervise the predicted mask M_{sc} .

3.4 Size-Aware Hungarian Matching (SAHM)

There are always inevitable noise points in the original annotations, such as missed labels, incorrect labels, and inaccurate positioning. Generally, a labeled point represents the entire head, and in most cases, the label is placed as close as possible to the center of the head. However, due to factors such as occlusion and viewpoint variations, label points may be deviated from the centers, as shown in Fig. 2. In fact, in our experiments, we observed that during training P2PNet [7] and PET [9], point location loss stays within a certain range before gradually decreasing. Therefore, we believe that introducing some tolerance for position errors is crucial when predicting the head points. Intrinsically, for smaller heads that are farther away, the positional error is smaller, while for larger heads in the near region, the positional error may be larger. To address this issue, we propose a scale-aware Hungarian matching module, which estimates head size in advance and adjusts the weight of the position error in the matching cost.

Head size estimation. Crowd images usually contain heads of different scales. Some density-based methods [15] dynamically select the variance of the Gaussian kernel σ by computing the KNN distance of head points. Similarly, for N head annotation points in an image, we calculate the average distance from each point P_i to its k nearest neighbors to estimate the head size S_i^{knn} :

$$S_i^{knn} = \alpha_{knn} * Dist_{knn}(P_i, k). \quad (2)$$

Meanwhile, inspired by the depth-adaptive density map in [20, 21], we estimate the head size S_i^{depth} based on the depth value of P_i . That is, the head size is inversely proportional with its depth for a given head:

$$S_i^{depth} = \frac{\alpha_{depth}}{D(P_i)}. \quad (3)$$

Here α_{knn} and α_{depth} represent the hyper-parameters for the KNN-based head size and the depth-based head size respectively. $D(P_i) \in (0,1)$ represents the depth of P_i . Then, we compare S_i^{knn} and S_i^{depth} , and adjust S_i^{knn} accordingly to estimate the final head size S_i as follows:

$$S_i = \begin{cases} S_i^{knn} * (1 + \delta), & \text{if } S_i^{depth} \geq S_i^{knn} * (1 + \delta), \\ S_i^{knn} / (1 + \delta), & \text{if } S_i^{depth} < S_i^{knn} / (1 + \delta), \\ S_i^{knn}, & \text{otherwise.} \end{cases} \quad (4)$$

The parameter δ is used to control the comparison between S_i^{knn} and S_i^{depth} . It should be noted that our approach focuses on estimating head size information at a coarse level, and we empirically set $\alpha_{knn} = 0.5$, $\alpha_{depth} = 3.0$, and $\delta = 0.5$ to obtain the estimated size for each annotated head point.

Cost balance of Hungarian match. In P2PNet [7] and PET [9], Hungarian matching is used to select the most appropriate set of candidate points from a global perspective. Specifically, given N annotation points $\{P_1, P_2, \dots, P_N\}$ and M predicted points $\{P'_1, P'_2, \dots, P'_M\}$, along with the classification scores of these M candidate points, a cost matrix C is constructed in Hungarian matching by calculating a weighted combination of point position error and classification score:

$$C_{ij} = w_p \cdot \text{dist}(P'_i, P_j) - \text{Score}_i. \quad (5)$$

In general, heads that are closer tend to be larger, while farther heads are smaller in an image. Using a fixed value for w_p might not properly balance the position errors and classification score of head points at different scales, which can affect the point selection in Hungarian matching. To better balance the impact of position errors and classification scores for different head sizes, we use the estimated head size S_j to compute the weight for the position error:

$$w_p = \frac{1}{w_{size} S_j}. \quad (6)$$

Here, w_{size} is the weight for head size. In this way, for smaller heads, the position error has a larger impact, while for larger heads, the weight is smaller.

3.5 Loss function

For the predicted scale mask M_{sc} , we compute the segmentation loss L_{seg} using cross-entropy (CE) loss as follows:

$$L_{seg} = CE(M_{sc}, M_{gt}). \quad (7)$$

Furthermore, we use the cross-entropy loss as the classification loss L_{cls} and the smooth L_1 loss as the point location loss L_{loc} . The final loss L is given by:

$$L = L_{cls} + \lambda_1 L_{loc} + \lambda_2 L_{seg}. \quad (8)$$

where λ_1 and λ_2 are the corresponding loss weights.

4 Experimental Results

4.1 Datasets

We evaluate our method on three publicly available datasets. **ShanghaiTech Part A dataset** contains 482 images collected from busy urban scenes, with high crowd densities. It is divided into 300 training images and 182 testing images. The scenes in this dataset are challenging due to significant scale variations and dense crowds. **ShanghaiTech Part B dataset** includes 716 images captured in outdoor environments with moderate crowd densities. It is split into 400 training images and 316 testing images.

This dataset has lower density compared to Part A, making it useful for testing performance under varied crowd conditions. **JHU-Crowd++ [32] dataset** contains 4,372 images with a wide range of crowd densities, from sparse to extremely dense. The dataset is partitioned into a training set (2,272 images), a validation set (500 images), and a test set (1,600 images). The large number of high-resolution images in this dataset necessitates additional pre-processing, such as patch cropping to facilitate model training and testing.

Table 1. The experimental results on the ShanghaiTech Part A, ShanghaiTech Part B, and JHU-Crowd++ datasets.

Method	SH Part A		SH Part B		JHU-Crowd++	
	MAE	MSE	MAE	MSE	MAE	MSE
GL [33]	61.3	95.4	7.3	11.7	59.9	259.5
CSRNet [22]	61.2	97.8	7.6	12.7	69.4	262.4
DM-Count [34]	59.7	95.7	7.4	11.8	-	-
MSGSA [35]	57.8	91.3	6.9	10.8	-	-
ChfL [36]	57.5	94.3	6.9	11.0	57.0	235.7
PESSNet [24]	57.3	95.9	6.4	9.9	70.91	256.33
FIDT [23]	57.0	103.4	6.9	11.8	66.6	253.6
CLTR [8]	56.9	95.2	6.5	10.6	59.5	240.6
ScSiNet [37]	55.77	90.23	6.79	10.95	-	-
ADSCNet [16]	55.4	97.7	6.4	11.3	-	-
SDA+DM [38]	55.0	92.7	-	-	59.3	248.9
Gramformer [39]	54.7	87.1	-	-	53.1	228.1
HMoDE [40]	54.4	87.4	6.2	9.8	55.7	<u>214.6</u>
PFDNet [14]	53.8	89.2	6.5	10.7	-	-
STNet [17]	52.85	83.64	6.25	10.30	<u>55.03</u>	202.69
P2PNet [7]	52.74	85.06	6.25	9.9	-	-
mPromt [41]	52.5	88.9	5.8	9.6	-	-
LDFNet [42]	51.3	80.7	6.8	11.6	-	-
PSL-Net [43]	49.9	<u>77.6</u>	5.8	<u>9.2</u>	-	-
PET [9]	<u>49.34</u>	78.77	6.19	9.69	58.5	238.0
Depth methods						
DAnet [28]	71.4	120.6	9.1	14.7	-	-
RDNet [20]	-	-	8.8	15.3	-	-
DPDNet [21]	66.6	120.3	7.9	12.4	-	-
DECCNet [30]	58.6	101.1	7.1	11.4	-	-
DigCrowd [29]	-	-	21.5	32.3	-	-
DACAN [27]	57.1	90.1	7.2	12.1	-	-
Ours	46.69	74.57	<u>5.85</u>	8.76	56.95	220.64

4.2 Evaluation Metrics

Two standard metrics, Mean Absolute Error (MAE) and Mean Squared Error (MSE), are used. They are defined as follows:

$$\begin{aligned} \text{MAE} &= \frac{1}{N} \sum_{i=1}^N |C_i^{\text{pred}} - C_i^{\text{gt}}|, \\ \text{MSE} &= \sqrt{\frac{1}{N} \sum_{i=1}^N (C_i^{\text{pred}} - C_i^{\text{gt}})^2}. \end{aligned} \quad (9)$$

C_i^{pred} and C_i^{gt} represent the predicted crowd count and ground truth count for the i -th image respectively, and N is the total number of test images.

4.3 Implementation Details

Our implementation is extended from PET [9], in which the VGG16 is adopted as the backbone for feature extraction. The Adam optimizer is adopted for training, with a learning rate of 1×10^{-5} for the backbone and 1×10^{-4} for the remaining components. The batch size is set as 8. Input images are randomly cropped into 256×256 patches, with zero-padding applied for smaller patches to meet the size requirement. For the JHU-Crowd++ dataset, the high-resolution training images are divided into 512×512 patches with 25% overlap, while testing images are split into non-overlapping 512×512 patches to ensure consistent evaluation.

During the data augmentation stage, we randomly select a scaling factor for both the width and height directions of the original image, denoted as w_{scale} and h_{scale} , respectively, where $w_{\text{scale}}, h_{\text{scale}} \in [0.8, 1.2]$. Meanwhile, for a head at point P_j in the image with an estimated head size S_j , we adjust the head size for the loss weight computation in Eq (6) as follows:

$$S'_j = S_j \cdot \sqrt{w_{\text{scale}} \cdot h_{\text{scale}}}. \quad (10)$$

In the Transformer Encoder, we integrate three DRWA modules with dilation rates of 4, 2, and 1, respectively. In the Transformer Decoder, two DRWA modules are used with dilation rates of 2 and 1. Thanks to the effectiveness of DRWA, we utilize a rectangular window of size 8×4 , which are only one-fourth of the window size 16×8 used in the PET [9]. During pre-processing, depth maps for each dataset are generated using the Depth Anything [31] Model. A threshold of $d_{th} = 0.2$ is applied to generate the ground truth scale mask M_{gt} .

For the weight adjustments of Hungarian matching, we use $w_{\text{size}} = 1.5$ for the ShanghaiTech datasets and $w_{\text{size}} = 2.0$ for the JHU-Crowd++ dataset. The loss weights are set as $\lambda_1 = 0.05$, and $\lambda_2 = 0.1$.

4.4 Results

As shown in Table 1, our method achieves SOTA results with MAE 46.69 and MSE 74.57 on the ShanghaiTech Part A dataset, which is 2.65 lower than the second-best MAE 49.34 of PET [9] and 3.03 lower than the second best MSE 77.6 of PSL-Net [42]. On the ShanghaiTech Part B, Our MSE 8.76 set the new SOTA, 0.44 lower than the second best 9.2 of PSL-Net [42]. Our MAE of 5.85 on the ShanghaiTech Part B is slightly lower than 5.8 of PSL-Net and mPrompt [40]. On the JHU-Crowd++ dataset, our MAE of 56.95 and MSE of 220.64 are higher than the SOTA MAE 53.1 of Gram-former [38] and MSE 202.69 of STNet [17]. We think the reason is JHU-Crowd++ dataset is very big, containing 2272 high resolution training images, therefore the training process is time consuming and we do not tune the super-parameters on this dataset.

Compared with other depth-based methods in the bottom panel of Table 1, our results are significantly better than all of them on the ShanghaiTech datasets. These results demonstrate the effectiveness of our method across diverse datasets with varying crowd densities and image complexities. Since no results of depth-based methods are reported on JHU-Crowd++ dataset, we cannot compare with them.

4.5 Ablation study

We conducted ablation experiments on the ShanghaiTech Part A dataset.

Impact of the Three Proposed Modules. We tested the impacts of the proposed three modules. For the DRWA module, we eliminated its effect by setting all dilation rates to 1. In the case of the SAHM module, we fixed the matching weights in the Hungarian algorithm to the values used in the PET method, specifically setting $w_p = 0.05$. For the DGSS module, we conducted experiments using only F_1 or F_2 feature map as the query in the decoder. As shown in Table 2, without SAHM module, MAE increased from 46.69 to 48.96, and MSE increased from 74.57 to 77.01. Removing DRWA module, MAE increased to 47.92, and MSE increased to 77.72. Removing DGSS module and using only $F_1(F_2)$ feature map, MAE deteriorated significantly to 52.87 (51.85), and MSE to 82.62(82.53). This demonstrates the effectiveness of the proposed three modules.

Table 2. Ablation experiments for the proposed three modules on the ShanghaiTech Part A dataset.

Method	MAE	MSE
DRWA+DGSS w/o SAHM	48.96	77.01
DGSS+SAHM w/o DRWA	47.92	77.72
DRWA+ F_1 +SAHM w/o DGSS	52.87	82.62
DRWA+ F_2 +SAHM w/o DGSS	51.85	82.53
DRWA+DGSS+SAHM	46.69	74.57

Effect of Depth Threshold. The choice of d_{th} affects the scale segmentation masks, which subsequently affects the selection of appropriate features from different network levels for accurate head position estimation. As shown in Table 3, we conducted

experiments on the ShanghaiTech Part A dataset with various d_{th} , and the best performance was achieved when $d_{th} = 0.20$.

Table 3. Effect of depth threshold d_{th} for scale segmentation.

d_{th}	MAE	MSE
0.15	49.42	75.76
0.20	46.69	74.57
0.25	48.88	79.84

Effect of Head Size Weight. The value of head size weight w_{size} in the SAHW module plays a critical role in determining the point-wise weighting during the matching process. We conducted experiments with different values of w_{size} . From Table 4, the best performance is achieved when w_{size} is set to 1.5. It is observed that head size weight w_{size} has a big effect on the results. On the JHU-Crowd++ dataset, we set w_{size} as 2.0 and did not test other settings. The results may be improved with other setting of w_{size} .

Table 4. Effect of head size weight w_{size} .

w_{size}	MAE	MSE
0.5	48.65	76.19
1.0	48.98	77.41
1.5	46.69	74.57
2.0	49.18	77.51

Effect of Loss Weights. We conduct extensive experiments on ShanghaiTech Part A by varying the values of λ_1 and λ_2 . As shown in Table 5, the best performance is achieved when $\lambda_1 = 0.05$, and $\lambda_2 = 0.1$.

Table 5. Effect of loss function weights.

(λ_1, λ_2)	MAE	MSE
(1.0, 1.0)	50.85	80.94
(1.0, 0.1)	48.17	77.37
(0.5, 0.1)	47.95	78.39
(0.05, 0.1)	46.69	74.57
(0.01, 0.1)	49.75	77.57

Table 6. Comparison of different depth estimation models.

Method	MAE	MSE
EVP[44]	48.25	77.58
ZoeDepth[45]	47.15	76.34
DepthAnything	46.69	74.57

Depth Estimation Model. We further evaluate the model’s compatibility with different monocular depth estimation models, including EVP [44] and ZoeDepth [45]. As reported in Table 6, DepthAnything model [31] performs best.

4.6 Visualization

We visualize the prediction results and compare with PET [9] in Fig. 7. In a close-range scene with larger head scales (the first row) and a distant and high-density crowd scene (the second row), our predicted counts are closer to the ground truth counts.

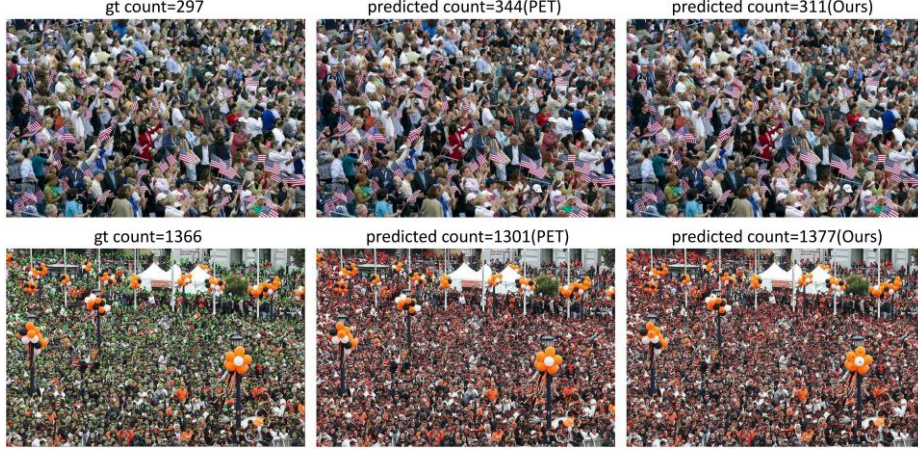


Fig. 6. Visualization and comparison of Crowd Counting Results Between Our Model and PET [9].

Table 7. Comparison of model parameters and inference time with point-based methods. The inference time is measured on a NVIDIA 3090 GPU with 1024×1024 input.

Method	Parameters (M)	Inference Time (s)
P2PNet	21.6	0.074
CLTR	43.4	0.107
PET	20.9	0.097
Ours	22.0	0.108

4.7 Model complexity

We compare the model parameters and inference time in Table 7. The numbers of other methods are from [9]. P2PNet [7] is a CNN-based method, which had the least inference time. CLTR [8] is a DETR-like method, in which the global attention is used, resulting in the most model parameters. PET [9] uses a transformer with window attention mechanism, which has much less parameters than CLTR. Our method uses dilated window attention and has slightly more parameters and inference time than PET [9].

5 Conclusion

In this paper, we propose the Depth Guided Scale-Aware Transformer (DGSAT) for the crowd counting task, which consists of three novel modules. Experimental results demonstrate the effectiveness of the proposed method.

There are some hyper-parameters in the proposed modules, for example, the depth threshold and the head size weight, which have to be set manually and empirically. In future work, we will explore how to implement the modules more efficiently. Due to the large number of heads in crowd images, we will also investigate how to optimize the Hungarian matching process.

References

1. Muhammad Asif Khan, Hamid Menouar, and Ridha Hamila. Revisiting crowd counting: State-of-the-art, trends, and future perspectives. *Image and Vision Computing*, 129:104597 (2023)
2. Lijia Deng, Qinghua Zhou, Shuihua Wang, Juan Manuel Górriz, and Yudong Zhang. Deep learning in crowd counting: A survey. *CAAI Transactions on Intelligence Technology*, 9(5), 1043–1077 (2024).
3. Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In *CVPR*, pp. 6133–6142, (2019)
4. Feng Dai, Hao Liu, Yike Ma, Xi Zhang, and Qiang Zhao. Dense scale network for crowd counting. In *Proceedings of the 2021 international conference on multimedia retrieval*, pp. 64–72, (2021)
5. Qingyu Song, Changan Wang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Jian Wu, and Jiayi Ma. To choose or to fuse? scale selection for crowd counting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2576–2583, (2021)
6. Changan Wang, Qingyu Song, Boshen Zhang, Yabiao Wang, Ying Tai, Xuyi Hu, Chengjie Wang, Jilin Li, Jiayi Ma, and Yang Wu. Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting. In *ICCV*, pp. 3234–3242, (2021)
7. Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *ICCV*, pp. 3365–3374, (2021)
8. Dingkan Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *ECCV*, pp. 38–54. Springer (2022)
9. Chengxin Liu, Hao Lu, Zhiguo Cao, and Tongliang Liu. Point-query quadtree for crowd counting, localization, and more. In *ICCV*, pp. 1676–1685, (2023)
10. I-Hsiang Chen, Wei-Ting Chen, Yu-Wei Liu, Ming-Hsuan Yang, Sy-Yen Kuo, et al. Improving point-based crowd counting and localization based on auxiliary point guidance. *arXiv preprint arXiv:2405.10589*, (2024)
11. Xinghao Ding, Fujin He, Zhirui Lin, Yu Wang, Huimin Guo, and Yue Huang. Crowd density estimation using fusion of multi-layer features. *IEEE Transactions on Intelligent Transportation Systems*, 22(8), 4776–4787 (2021)
12. Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In Vittorio Ferrari, Martial Hebert, Cristian

- Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pp. 757–773, Cham, Springer International Publishing (2018)
13. Xingjiao Wu, Yingbin Zheng, Hao Ye, Wenxin Hu, Tianlong Ma, Jing Yang, and Liang He. Counting crowds with varying densities via adaptive scenario discovery framework. *Neuro-computing*, 397, 127–138 (2020)
 14. Zhaoyi Yan, Ruimao Zhang, Hongzhi Zhang, Qingfu Zhang, and Wangmeng Zuo. Crowd counting via perspective-guided fractional-dilation convolution. *IEEE Transactions on Multimedia*, 24, 2633–2647 (2022)
 15. Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pp. 1091–1100, (2018)
 16. Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4593–4602, (2020)
 17. Mingjie Wang, Hao Cai, Xian-Feng Han, Jun Zhou, and Minglun Gong. Stnet: Scale tree network with multi-level auxiliator for crowd counting. *IEEE Transactions on Multimedia*, 25, 2074–2084 (2023)
 18. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, (2021)
 19. Dingkan Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. Transcrowd: weakly-supervised crowd counting with transformers. *Science China Information Sciences*, 65(6), 160104 (2022).
 20. Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao. Density map regression guided detection network for rgb-d crowd counting and localization. In *CVPR*, pp. 1821–1830, (2019)
 21. Dongze Lian, Xianing Chen, Jing Li, Weixin Luo, and Shenghua Gao. Locating and counting heads in crowds with a depth prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 9056–9072 (2022)
 22. Zhi-Qi Cheng, Qi Dai, Hong Li, Jingkuan Song, Xiao Wu, and Alexander G Hauptmann. Rethinking spatial invariance of convolutional networks for object counting. In *CVPR*, pp. 19638–19648, (2022)
 23. Dingkan Liang, Wei Xu, Yingying Zhu, and Yu Zhou. Focal inverse distance transform maps for crowd localization. *IEEE Transactions on Multimedia*, 25, 6040–6052 (2023)
 24. Jun Yi, Yiran Pang, Wei Zhou, Meng Zhao, and Fujian Zheng. A perspective-embedded scale-selection network for crowd counting in public transportation. *IEEE Transactions on Intelligent Transportation Systems*, 25(5), 3420–3432 (2024)
 25. Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv pre-print arXiv:2209.15001*, (2022)
 26. Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Andy J Ma, Yaowei Wang, and Wei-Shi Zheng. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Transactions on Multimedia*, 25, 8906–8919 (2023)
 27. Hao Chen, Zhan Li, Bir Bhanu, Dongping Lu, and Xuming Han. A depth-guided attention strategy for crowd counting. In Lazaros Iliadis, Antonios Papaleonidas, Plamen Angelov, and Chrisina Jayne, editors, *Artificial Neural Networks and Machine Learning – ICANN 2023*, pp. 25–37, Cham, Springer Nature Switzerland (2023)
 28. Van-Su Huynh, Vu-Hoang Tran, and Ching-Chun Huang. Danet: Depth-aware network for crowd counting. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3001–3005, (2019)

29. Mingliang Xu, Zhaoyang Ge, Xiaoheng Jiang, Gaoge Cui, Pei Lv, Bing Zhou, and Changsheng Xu. Depth information guided crowd counting for complex crowd scenes. *Pattern Recognition Letters*, 125, 563–569 (2019)
30. Shuo-Diao Yang, Hung-Ting Su, Winston H Hsu, and Wen-Chin Chen. Deccnnet: Depth enhanced crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4521–4530, (2019)
31. Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, (2024)
32. Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE trans on PAMI*, 44(5), 2594–2609 (2020)
33. Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *CVPR*, pp. 1974–1983, (2021)
34. Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. *Advances in neural information processing systems*, 33, pp.1595–1607, (2020)
35. Yange Sun, Meng Li, Huaping Guo, and Li Zhang. Msgsa: Multi-scale guided self-attention network for crowd counting. *Electronics*, 12(12), (2023)
36. Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *CVPR*, pp. 19618–19627, (2022)
37. Mingjie Wang, Hao Cai, Jun Zhou, and Minglun Gong. Interlayer and intralayer scale aggregation for scale-invariant crowd counting. *Neurocomputing*, 441, 128–137 (2021)
38. Zhiheng Ma, Xiaopeng Hong, Xing Wei, Yunfeng Qiu, and Yihong Gong. Towards a universal model for cross-dataset crowd counting. In *ICCV*, pp. 3205–3214, (2021)
39. Hui Lin, Zhiheng Ma, Xiaopeng Hong, Qinnan Shangguan, and Deyu Meng. Gram-former: Learning crowd counting via graph-modulated transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3395–3403, (2024)
40. Zhipeng Du, Miaojing Shi, Jiankang Deng, and Stefanos Zafeiriou. Redesigning multi-scale neural network for crowd counting. *IEEE Transactions on Image Processing*, (2023)
41. Mingyue Guo, Li Yuan, Zhaoyi Yan, Binghui Chen, Yaowei Wang, and Qixiang Ye. Regressor-segmenter mutual prompt learning for crowd counting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28380–28389, (2024)
42. Yuehai Chen, Qingzhong Wang, Jing Yang, Badong Chen, Haoyi Xiong, and Shaoyi Du. Learning discriminative features for crowd counting. *IEEE Transactions on Image Processing*, 33, 3749–3764 (2024)
43. Jihye Ryu and Kwangho Song. Crowd counting and individual localization using pseudo square label. *IEEE Access*, (2024)
44. Mykola Lavreniuk, Shariq Farooq Bhat, Matthias Müller, and Peter Wonka. Evp: Enhanced visual perception using inverse multi-attentive feature refinement and regularized image-text alignment. *arXiv:2312.08548*, (2023)
45. Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv:2302.12288*, (2023)