CrowdFormer: An Overlap Patching Vision Transformer for Top-Down Crowd Counting

Shangpeng Yang[2] , **Weiyu Guo**[1,*] , **Yuheng Ren**[2]
[1]Information School, Central University of Finance and Economics, Beijing, China
[2]Watrix Technology Co. LTD., Beijing, China
shaopeng.yang@watrix.ai, weiyu.guo@cufe.edu.cn, yuheng.ren@watrix.ai

Abstract

Crowd counting methods typically predict a density map as an intermediate representation of counting, and achieve good performance. However, due to the perspective phenomenon, there is a scale variation in real scenes, which causes the density map-based methods suffer from a severe scene generalization problem because only a limited number of scales are fitted in density map prediction and generation. To address this issue, we propose a novel vision transformer network, i.e., CrowdFormer, and a density kernels fusion framework for more accurate density map estimation and generation, respectively. Thereafter, we incorporate these two innovations into an adaptive learning system, which can take both the annotation dot map and original image as input, and jointly learns the density map estimator and generator within an end-to-end framework. The experimental results demonstrate that the proposed model achieves the state-of-the-art in the terms of MAE and MSE (e.g., it achieved a MAE of 67.1 and MSE of 301.6 on NWPU-Crowd dataset.), and confirm the effectiveness of the proposed two designs. The code is https://github.com/special-yang/Top_Down-CrowdCounting.

1 Introduction

Crowd counting is an essential topic in a variety of applications such as public safety, activity recognition, and transportation management, aiming to estimate the number of people in a crowd image. Current methods[Wan *et al.*, 2020; Jia *et al.*, 2021; Boyu *et al.*, 2020] obtained excellent progress by utilizing the CNNs to regress the corresponding density maps of the input images, where the summed value in a density map indicates the total counting of people in the crowd image. The density map is an intermediate representation of the crowd, which groundtruth is typically generated from the original image by placing a density kernel on each person's dot annotation.

Due to the perspective phenomenon in real crowd scenes, there is a large-scale variation on different targets, which necessitates a crowd counting method that can deal with various scales. However, the current density map based crowd counting methods remain the following drawbacks: 1) Due to the limited and fixed receptive fields of pure CNNs, which cannot fit a wider range of continuous scale variation of targets, their density map estimators typically only deal with discrete scales. 2) For an efficient counting context information extraction, the dense or far-focus crowd scenarios are better suitable to a small density kernel, whereas sparse or close-focus crowd scenarios adapt to a large-scaled kernel. However, the groundtruth of density maps are typically generated by placing a hand-crafted density kernel on the corresponding dot annotations, which cannot fit to the scale change of the targets and scenarios. Based on the preceding observations, it is preferable to exploit a comprehensive solution for coping the continuous scale variation of crowds in both density map estimation and generation.

A human can estimate the approximate number of people in a crowd image quickly using a global to local visual perception mechanism, without the scale and receptive field problems. People using the Top-Down visual perception mechanism typically scan a crowd image with a global receptive filed, and then estimate the degree of congestion of crowd regions based on the prior knowledge from the global scanning. Furthermore, Transformer architectures can provide a global to local receptive field[Zheng *et al.*, 2021]. Inspired by the human's Top-Down visual perception mechanism and recent research progress about the Transformer, we propose a Crowd Counting Transformer network, namely, CrowdFormer, which models the human's Top-Down visual perception mechanism based on a series of Overlap Patching Transformer (OPT) blocks. An OPT block first learns global prior knowledge about a crowd image from an overlap visual patch sequence before focusing on analysing local crowd regions. In the OPT block, we can obtain a crowd counting from global to local, and encode the relative spatial position of visual tokens rather than absolute position, which is adopted in the standard vision transformer[Alexey *et al.*, 2021] and may result in the predictive model losing the translation and rotation invariance.

Moreover, to generate more accurate groundtruth of density maps considering various target scales, we propose a multiple density kernels fusion-based density map generator (KFMG), which can fit targets of various scales by fusing

[*] Corresponding author

multiple density kernels. Finally, we incorporate the density map generator and the estimator into an adaptive learning system, which takes both the dot annotation map and original image as input, and jointly learns the density map estimator and generator within an end-to-end framework.

In a nutshell, the main contributions of this paper are as follows:

- We propose to model the human's Top-Down visual perception mechanism in crowd counting by using the Transformer architecture. To the best of our knowledge, this is the first work to utilize Transformer architecture network for regressing density maps in crowd counting.
- A learnable density map generator, which takes various target scales into account, is proposed to generate more accurate density map groundtruths by fusing multiple density kernels.
- Extensive experiments are carried out to validate the proposed method, and demonstrate that our method is the-state-of-the-art from multiple perspectives.

2 Related Work

Crowd counting is a fundamental task in computing vision. In the early phase, people usually treat the crowd counting as target localization tasks. For example, Min Li et al.[Min *et al.*, 2008] used heads and shoulders localization to construct crowd counting. However, in severely occluded scenarios, such paradigms will be invalid. To cope with the densely crowded scene, two types of regression-based methods were proposed: global regression and density map regression. The global regression-based methods[Chan *et al.*, 2008; Chattopadhyay *et al.*, 2017; Yifan *et al.*, 2020] typically estimate the final count of people directly from images, while the density map regression-based approaches[Yuhong *et al.*, 2018; Shuai *et al.*, 2020; Jia *et al.*, 2020] first predicts a density map, which is then summed to obtain the final count. Because density maps contain more spatial context information, the performance is usually superior to that of global regression-based methods. Our work falls under the category of density map regression-based methods. As a result, we primarily examine two types of density map regression-based methods in recent literature. Moreover, since we introduce the Transformer architecture into the crowd counting task, existing Transformer approaches are also discussed.

2.1 Density Map Ground-truth Generation

The use of density maps as the supervised information is a common choice among most of recent crowd counting methods. Traditional work[Lempitsky *et al.*, 2010] typically convolve the crowd image by placing a fixed bandwidth Gaussian kernel on the dot annotation map to obtain a density map groundtruth. However, due to target scale variation, the Gaussian kernel with fixed bandwidth and parameters cannot contribute to all variation in existing images.

To cope with the scale variation on targets, some work was proposed to fit the scale variation by adjusting the bandwidth or parameters of the Gaussian kernel manually according to the scene perspective or crowdedness. For example, Idrees et al.[Haroon *et al.*, 2018] fuse multiple density maps which were generated by multiple Gaussian kernels with different bandwidths. Wan et al.[Wan and Chan, 2019; Wan *et al.*, 2020] propose using a CNN-based network as the density map generator. The parameters of the Gaussian kernel are changed in this way to make them learnable. Although many compelling methods for improving the quality of density map generation have been proposed, the problem of scale variation on targets remains unsolved due to the hand-crafted settings lacking generalization. In contrast to these previous work that requires manually setting the parameters or bandwidth of Gaussian kernel, we propose to adaptively fuse multiple density kernels, and jointly learn the density map generator with the estimator in an end-to-end manner.

2.2 Multiscale Density Map Estimation

To deal with a large-scale variation on targets, some literature proposed using multiscale features or multicontext information during the feature extraction phase. As a result, using the feature pyramid network (FPN)[Tsung-Yi *et al.*, 2017] to fuse multi-scale information is a common choice among recent crowd counting methods. For example, Chen et al.[Chen *et al.*, 2021] utilized FPN to extract multiscale features with different receptive fields, whereas a pyramid region awareness loss[Qingyu *et al.*, 2021] is utilized to recursively searches the most over-estimated sub-regions.

Although multiscale features and multicontext information about targets have been focused on in recent literature, these multiscale learning-based methods still fail to cope with the continuous scale variation on targets. In contrast to previous work that only consider a few scales, we present CrowdFormer, a novel Transformer network, that can fit targets with continuous scales by modeling the human Top-Down visual perception mechanism.

2.3 Transformers in Crowd Counting

CNNs are regarded as a hierarchical ensemble of local features with different reception fields. Unfortunately, most CNNs excel at extracting local features but struggle to capture global cues. Transformer[Ashish *et al.*, 2017], which was proposed in the field of natural language processing to capture the long-term dependencies between input and output, was recently introduced to vision tasks. Due to fusing local and global features, Transformers promote the performance of many vision tasks significantly. For example, Dosovitskiy et al.[Alexey *et al.*, 2021], constructed a Transformer network for image classification, and achieved excellent results compared to CNN-based models. While, Nicolas et al.[Nicolas *et al.*, 2020], utilized Transformer structures to augment a standard CNN network for improving the performance on object detection. Moreover, Liang et al.[Dingkang *et al.*, 2021] proposed TransCrowd, which involves a Transformer network into the weakly-supervised crowd counting task, and achieves good performance. However, because the final counts are directly regressed the TransCrowd fails to learn the spatial and contextual information of scenes. In contrast to TransCrowd, we propose a novel cutting-edge Transformer framework for density map estimation that takes efficiency, accuracy, and robustness into consideration.

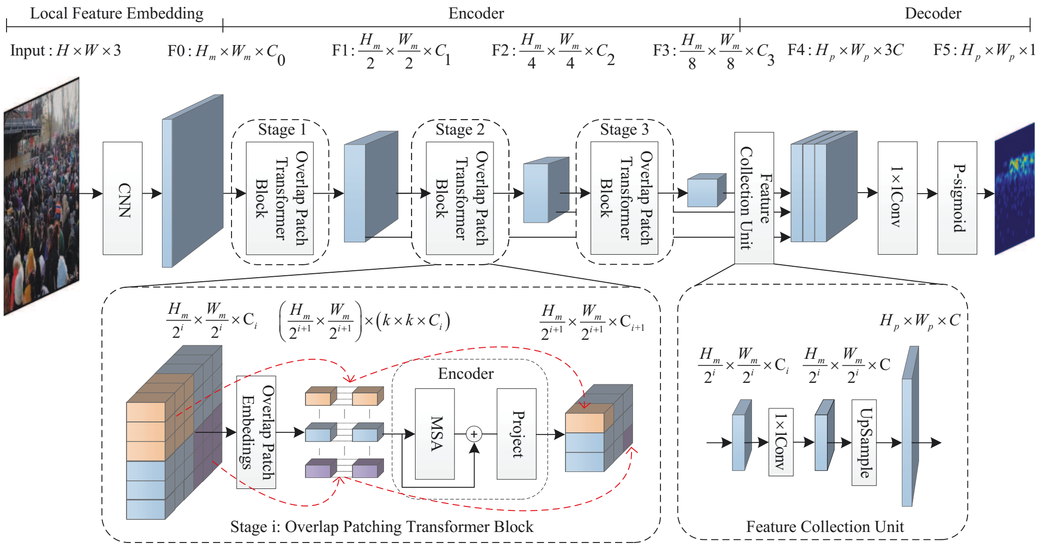
Local Feature Embedding
Encoder
Decoder
Input : $H \times W \times 3$
F0 : $H_m \times W_m \times C_0$
F1 : $\frac{H_m}{2} \times \frac{W_m}{2} \times C_1$
F2 : $\frac{H_m}{4} \times \frac{W_m}{4} \times C_2$
F3 : $\frac{H_m}{8} \times \frac{W_m}{8} \times C_3$
F4 : $H_p \times W_p \times 3C$
F5 : $H_p \times W_p \times 1$
CNN
Stage 1
Stage 2
Stage 3
Overlap Patch Transformer Block
Feature Collection Unit
1×1Conv
P-sigmoid
$\frac{H_m}{2^i} \times \frac{W_m}{2^i} \times C_i$
$\left(\frac{H_m}{2^{i+1}} \times \frac{W_m}{2^{i+1}}\right) \times (k \times k \times C_i)$
$\frac{H_m}{2^{i+1}} \times \frac{W_m}{2^{i+1}} \times C_{i+1}$
Overlap Patch Embedings
Encoder
MSA
Project
Stage i: Overlap Patching Transformer Block
$H_p \times W_p \times C$
$\frac{H_m}{2^i} \times \frac{W_m}{2^i} \times C_i$
$\frac{H_m}{2^i} \times \frac{W_m}{2^i} \times C$
1×1Conv
UpSample
Feature Collection Unit

Figure 1: Overview of CrowdFormer. The CrowdFormer framework mainly consists of three modules: a local feature embedding module for efficiently learning abstract and low-resolution feature maps from large input images; a vision Transformer-based encoder for extracting coarse and fine features; and a lightweight decoder for directly fusing these multilevel features and generating the predictive density map.

3 CrowdFormer

In the real scene, there is a large-scale variation on different targets due to the perspective phenomenon. Recent work typically only can fit a few scales. However, Humans, on the other hand, are usually not affected by such scale variation, since we can take the global context information of the crowd scene into account with a global receptive field. In light of this, we propose CrowdFormer, which is capable of modeling the human's Top-Down visual perception mechanism in crowd counting task.

The CrowdFormer cascades CNN and visual Transformer structures for local crowd region enhancement and global context information capturing. An overview of CrowdFormer is depicted in Figure 1, which is made up of three major modules, i.e., a local feature embedding module, a Transformer structure-based encoder, and a lightweight decoder. To begin, the local feature embedding module extracts local features about crowd counting, and embeds them into the given low-resolution representations using a CNN structure inherited from the YOLOv5 backbone. Then, the encoder, which is stacked by our proposed Transformer blocks to generate feature maps of different resolution, is used to fuse the local crowd features and the global information of crowd scene. Finally, the lightweight decoder combines the multi-scale feature maps to predict a density map.

3.1 Local Feature Embedding

Considering that self-attention-based Transformer when performed across n entities requires $O(n^2)$ memory and computation, which is memory and computation consuming. While CNN can extract local but compact feature representation from the inputs efficiently. We first implement a CNN feature extractor to obtain a compact feature representation from the inputs for computing resource saving and local information refinement.

Specifically, the structure of local feature extractor we adopted in the CrowdFormer is inherited from the backbone of YOLOv5, and used for 1/4 downsampling feature extraction. In this way, we can obtain a good fitting ability with a small computing cost, as well as alleviate the effective information loss in the downsampling process. In practice, we directly use the part of pretrained backbone of YOLOv5 as the local feature extractor, because the tinny object detection and crowd counting have similarities in terms of task properties and their domain knowledge may can be transferred to each other.

3.2 Overlap Patching Transformer Block

The human could focus on the region of interest in the image, based on the goal of current task and global prior knowledge. Similarly, we boost the model attention on crowd regions by mining global context information. As a result, we propose a Transformer-based encoder to model the human Top-Down visual perception mechanism.

The encoder is stacked by Overlap Patching Transformer (OPT) blocks, which can mining the global prior knowledge of the crowd scene to reinforce the local crowd region feature from coarse to fine-grained. As shown in Figure.1, given the feature maps with a resolution of $H_m \times W_m$, our encoder performs i-th OPT block to obtain feature maps $F_i \in \mathbb{R}^{\frac{H_m}{2^i} \times \frac{W_m}{2^i} \times C_i}$, where $i \in \{1, 2, 3\}$. An OPT composed of two sequential parts: an overlap patching layer and a self-attention based feature encoder.

The overlap patching layer is to resolve the inputted 3D feature maps into a sequence of 2D patches $x_i \in \mathbb{R}^{N \times D}$ like the stander vision transformer[Alexey *et al.*, 2021], where $N = \frac{H_i}{r_i} \times \frac{W_i}{r_i}$, $D = k_i \times k_i \times C_i$, $k_i \times k_i$ is the resolution of the patches and r_i is the stride of the sliding window. Instead of splitting inputs into non-overlapping patches in manner of ViT[Alexey *et al.*, 2021], we use a $k_i \times k_i$ slide window with stride r_i ($r_i < k_i$ and $r_i = 2$) on inputted feature maps to obtain overlapped 2D feature patch sequence, which patches maintain the relative position relation, while traditional transformers patch an input without overlapping, and may lost the relative position information between patches.

Transformer encoder utilizes a multi-head self-attention module (MSA) to learn the global prior knowledge about crowd sense from the patches x_i in multiple feature subspace. Then, a learnable projector is used to project patches into their relative spatial positions, which is shown as in the overlap patch transformer block of Figure 1. Formally, the feature encoder can be denoted as:

$$\begin{aligned} x'_i &= MSA(x_i) + x_i \\ F_{i+1} &= fold(x'_i\ , W_i) \end{aligned} \tag{1}$$

where the MSA first splits the query, key, and value matrices of the standard self-attention into h ploids and performs attention functions of these ploids in parallel:

$$\begin{aligned} &[Q_{i,p}, K_{i,p}, V_{i,p}] = x_i \cdot [W_{Q_{i,p}}, W_{K_{i,p}}, W_{V_{i,p}}] \\ &x'_{i,p} = softmax(\frac{Q_{i,p} \cdot K_{i,p}^T}{\sqrt{d}}) \cdot V_{i,p} \\ &x'_i = cat(x'_{i,1}, .., x'_{i,p}, ..., x'_{i,h}) \end{aligned} \tag{2}$$

where $Q_{i,p}, K_{i,p}, V_{i,p} \in \mathbb{R}^{N \times \frac{D}{h}}$ are the query, key, value, and input matrices of the p-th ploid self-attention, and $W_{Q_{i,p}}, W_{K_{i,p}}, W_{V_{i,p}} \in \mathbb{R}^{\frac{D}{h} \times \frac{D}{h}}$ are the corresponding mapping matrices. The $cat(*)$ concatenates the outputs of h self-attentions together.

Specifically, $fold(*\ ,*)$ is a compound operator to keep the relative position of patches unchanged, and fold the patch sequence x'_i into feature maps $F_i \in \mathbb{R}^{\frac{H_i}{r_i} \times \frac{W_i}{r_i} \times C_i}$. It first reshapes the learnable project tensor $W_i \in \mathbb{R}^{C_i \times C_{i-1} \times k_i \times k_i}$ to be the $\hat{W}_i \in \mathbb{R}^{C_i \times D}$, and then performs a dot product between x'_i and $\hat{W}_i^T$. Finally, the result of dot product is fold into F_i. With the weights sharing and overlap patching, similar to convolution network, the relative spatial position information of patches can be encoded into their feature representation. Comparing with adding the absolute position embeddings with feature vectors of patch sequence used in the standard vision Transformer[Alexey *et al.*, 2021], proposed method may result in improved generalization performance in the terms of translation and rotation. It should be noted that, recent work[Ze *et al.*, 2021] shows that encoding the absolute position of vision tokens may led to losing the translation and rotation invariance, which are vital to the generalization performance of visual models.

3.3 Density Map Estimation

To predict the final density map, we propose a lightweight decoder. As shown in the decoder phase of Figure.1, each feature F_i outputted by the corresponding OPT block is sent into a feature collection unit (CFU), which utilizes a 1×1 convolution layer to fuse the F_i to be the $\hat{F}_i \in \mathbb{R}^{\frac{H_m}{r_i} \times \frac{W_m}{r_i} \times C}$, and then upsamples each $\hat{F}_i$ with bilinear interpolation into the given resolution $H_p \times W_p$. Then, all features $\hat{F}_i$ are concatenated to be the feature $F \in \mathbb{R}^{H_p \times W_p \times 3C}$, where $C = 256$ is a hyper-parameter, and a 1×1 convolution layer following a novel activation function, i.e., P-Sigmoid, is used to fuse the different level features and generate the density map prediction.

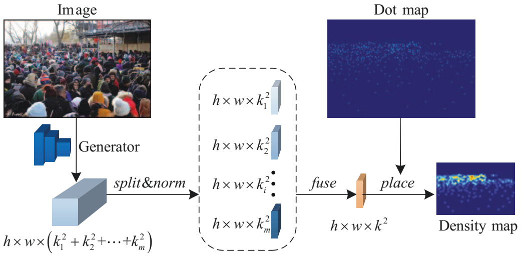
Image
Dot map
Generator
split&norm
$h \times w \times k_1^2$
$h \times w \times k_2^2$
$h \times w \times k_i^2$
$h \times w \times k_m^2$
fuse
place
$h \times w \times (k_1^2 + k_2^2 + \cdots + k_m^2)$
$h \times w \times k^2$
Density map

Figure 2: Multiple kernels fusion-based density map generation.

Specifically, P-Sigmoid can be defined as:

$$x_{out} = \frac{a}{1 + e^{-x}} \tag{3}$$

where a is a learnable parameter. It should note that a point on density map may express more than one person. Usually used activation functions in deep learning, e.g., $tanh$, $Sigmoid$ and PReLU, are unsuitable for this task due to their ranges. Therefore, most of existing methods adopt the absolute function to be the activation function for generating the final densisy map. In comparison to the absolute function, the scale of P-Sigmoid values are close to the absolute function due to the learnable parameter a, as well as the gradients are smoother due to having the attributes of $Sigmoid$. The proposed P-Sigmoid is better suited to crowd counting.

4 Density Kernels Learning and Fusion

Previous related work mainly focuses on the density map estimation but ignores the generation of density map label. They usually generate a density map label by using a hand-crafted Gaussian kernel to convolve the dot annotation map, where the spatial location of each annotated person takes the value 1, (and 0 otherwise). In this way, these work failed to fit the scale change of the targets and scenarios. As a result, we propose a multiple density kernels fusion based density map generator, i.e.,KFMG, to generate density map label taking multiple scales content information into consideration. The pipeline is shown in Figure 2. We first generate multiple adaptive density kernels with different sizes from the original image, and then fuse those density kernels to be one fused kernel. Finally, the fused kernel is used to convolve the dot annotation map and generate the density map label.

Given an input image I, we construct a CNN to generate multiple adaptive density kernels $K_1, K_2, .., K_m$, and fuse them to be one adaptive density kernel $\hat{K}$. Formally, the generating process can be denoted as:

$$\begin{aligned} &\mathrm{K} = \Omega(I) \\ &K_1, K_2, .., K_m = split(\mathrm{K}; [k_1^2, k_2^2, .., k_m^2]) \\ &\hat{K} = \sum_{i=1}^{m} Normalize(K_i)/m \end{aligned} \tag{4}$$

where $\Omega(*)$ denotes a CNN, which first downsamples I to be the feature maps $\mathrm{K} \in \mathbb{R}^{H_p \times W_p \times C_K}$ with the given resolution of $H_p \times W_p$. $C_K = \sum_{i=1}^{m} k_i^2$ is the number of channels in K, and k_i is the size of the i-th adaptive density kernel. Then, $split(\text{K}; [k_1^2, k_2^2, .., k_m^2])$ is to divide K into m groups on the channel dimension, and each group K_i can be treated as an adaptive density kernel with the given size of $k_i \times k_i$. Finally, these adaptive density kernels are aligned along with their center points, and fused to be one adaptive density kernel $\hat{K}$ with the size of $\hat{k} \times \hat{k}$, where $\hat{k} = \max(k_1, k_2, ..., k_m)$.

Let $D = \{p_j\}_{j=1}^{N}$ be the set of N annotated 2D coordinates of the persons in dot annotation map, $p_j = (x_j, y_j)$ in the image I. For each annotation position p_j, we retrieve the corresponding kernel map $\hat{K}_{p_j} \in \mathbb{R}^{\hat{k} \times \hat{k}}$ from $\hat{K}$, and then normalize it to sum to 1, resulting in the location-specific kernel $\hat{k}_{p_j} = \hat{K}_{p_j} / sum(\hat{K}_{p_j})$. Finally, the density map is generated by placing the location-specific kernel maps on the corresponding positions in the density map,

$$M(p) = \sum_{j=1}^{N} \hat{k}_{p_j}(p - p_j) \tag{5}$$

where the indexing of $\hat{k}_{p_j}$ is on $p \in (-r, ..., r) \times (-r, ..., r)$, and $r = (\hat{k} - 1)/2$. It is important to note that the equation (5) is analogous to placing a Gaussians kernel on each dot annotation to generate a traditional density map, except that the multiple kernels are learned, and could be different for each position and each image. The learned density map M is then used as the groundtruth to train our density map estimator.

5 Experiments

In this section, we conduct experiments evaluating our proposed CrowdFormer and density map generation framework. First, we go over the datasets, training, and evaluation metrics in detail. Second, we compare our method with other approaches. Finally, we ablate the key design elements of our framework.

5.1 Experimental Setup

In this experiment, three real-world datasets, i.e., NWPU-Crowd[Qi *et al.*, 2021], UCF-QNRF[Haroon *et al.*, 2018], and ShanghaiTech[Yingying *et al.*, 2016], are used for evaluation in our experiments.

In model training, the CNN based local feature embedding are initialized from the first three blocks of pre-trained YOLOv5, and the rest parts of network are randomly initialized by a Gaussian distribution with the mean of 0 and the standard deviation of 0.01. The scales of adaptive density kernels we adopted for generating the density map groundtruth are 3 × 3 and 5 × 5, respectively. For parameters training, an Adam optimizer are employed 900 epochs with a cosine decay learning rate scheduler and 10 epochs of linear warm-up. The initial learning rate and weight decay are set to 1e-5 and 1e-4, respectively. The training batch size is set to 12, and data augmentations such as random-cropping of raw input with size 512 × 512, random horizontal flipping and color jittering are adopted.

During the training, we incorporated the proposed density map estimator and generator into an adaptive system, which takes the both annotation dot map and original image as input, and jointly learns the estimator and generator by pixel-level L_2 loss between the estimation and the generation of density maps within an end-to-end framework.

5.2 Evaluation and Analysis

On four real-world datasets, we compare our proposed density map prediction framework to previous state-of-the-art methods to assess overall counting performance. Following the previous work, we adopt Mean Absolute Error (MAE) and Mean Squared Error (MSE) as the criterion to evaluate the counting performance. According to the results shown in Table1, comprehensively compared with previous methods, the performance of our method is the state-of-the-art, despite being challenged by other approaches in the highly congested scene, i.e., ShTech A.

Specifically, the TransCrowd[Dingkang *et al.*, 2021], which also use the Transformer architecture for crowd counting task, appears to be inadequate to other methods. However, due to CrowdFormer, which encodes relative spatial position of visual tokens rather than absolute position, our approach outperforms TransCrowd on a variety of benchmarks. Furthermore, on the largest dataset, i.e., NWPU, our method achieves the best performance, which improves the state-of-the-arts by 13.3% on MAE and 12.8% on MSE. This result validates the effectiveness of the proposed method in real-world application. Finally, our method outperforms the dot counting regression methods, such as NoiseCC[Jia *et al.*, 2020], DM-count[Boyu *et al.*, 2020], BM-count[Liu *et al.*, 2021], GLoss[Jia *et al.*, 2021] and P2PNet[Song *et al.*, 2021], on most of the datasets, because our density map generation method can utilize more context information about crowd counting during both the training and testing phases, and allowing us to deal with a wide range of complex crowd scenes. Moreover, comparing with the recent density map and multiscale methods, such as KDMG[Wan *et al.*, 2020], DKPNet[Chen *et al.*, 2021], SASNet[Qingyu *et al.*, 2021] and S3[Lin *et al.*, 2021], our method also can obtain significant advantage on most of the datasets, because the CrowdFormer can obtain a crowd counting in manner of global to local.

5.3 Ablation Studies

Ablation on the structure of CrowdFormer: We examine the effectiveness of the CrowdFormer structure. As the results shown in Table 2, when we use 3 × 3 convolution layers to replace all OPT blocks in CrowdFormer, the performance of crowd counting, i.e., Non_OPT, degrades significantly when compared to the performance of CrowdFormer. Moreover, when we use an OPT block to replace the local feature embedding module of CrowdFormer, the performance, i.e., Pure_OPT, can approach the performance of CrowdFormer, but a huge computation cost. These results demonstrate that the effectiveness of the CrowdFormer structure on context information extraction about crowd regions.

Ablation on activation function: We substitute a common used absolute function (Abs) for the P-Sigmoid to determine the effect of the proposed P-Sigmoid activation function. In detail, we use model training and testing to demonstrate the

Methods	Venue	Features				NWPU		UCF-QNRF		ShTech A		ShTech B	
		dr	dmg	ms	vt	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
NoiseCC	NeurIPS 20	✓	×	×	×	96.9	534.2	85.8	150.6	61.9	99.6	7.4	11.3
DM-count	NeurIPS 20	✓	×	×	×	88.4	357.6	85.6	148.3	59.7	95.7	7.4	11.8
BM-count	IJCAI 21	✓	×	×	×	83.4	358.4	81.2	138.6	57.3	90.7	7.4	11.8
GLoss	CVPR 21	✓	×	×	×	79.3	346.1	84.3	147.5	61.3	95.4	7.3	11.4
P2PNet	ICCV 21	✓	×	×	×	77.4	362.0	85.3	154.5	**52.7**	**85.1**	6.2	9.9
KDMG	PAMI 20	×	✓	✓	×	100.5	415.5	99.5	173.0	63.8	99.2	7.8	12.7
DKPNet	ICCV 21	×	✓	✓	×	74.5	327.4	81.4	147.2	55.6	91.0	6.6	10.9
SASNet	AAAI 21	×	✓	✓	×	--	--	85.2	147.3	53.6	88.4	6.4	9.9
S3	IJCAI 21	✓	✓	×	×	83.5	346.9	80.6	139.8	57.0	96.0	6.3	10.6
TransCrowd	--	✓	×	×	✓	117.7	451.0	97.2	168.5	66.1	105.1	9.3	16.1
CrowdFormer	--	×	✓	✓	✓	**67.1**	**301.6**	**78.8**	**136.1**	56.9	97.4	**5.7**	**9.6**

Table 1: Comparison with the state-of-the-art methods on benchmark datasets, where “dr”, “dmg”, “ms”, and “vt” indicate whether the methods belong or not to the categories of dot counting regression, density map generation, multiscale fusion, and using vision transformer architecture, respectively. In this experiment, we fuse two density kernels with sizes of 3×3 and 5×5 for generating the density map labels.

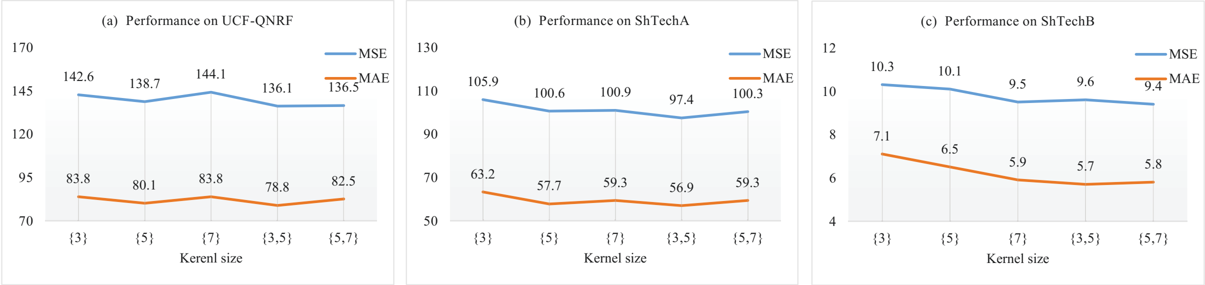

Figure 3: Testing performance on different datasets w.r.t. using different sizes of density kernels for density map generation.

Models	MSE	MAE
Non_OPT	152.3	90.2
Pure_OPT	138.3	81.1
CrowdFormer	**136.1**	**78.8**

Table 2: Comparison of different model structures on the UCF-QNRF dataset.

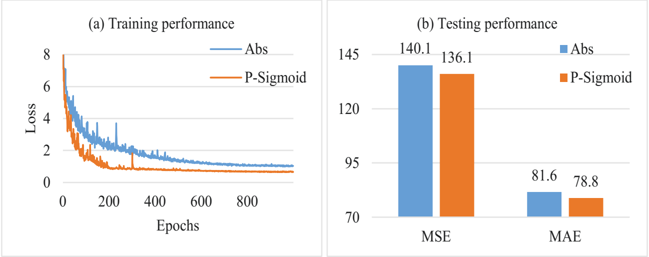

Figure 4: Training and testing performance w.r.t. using different activation functions for density map generation.

performance gain. As shown in Figure 4(a), in the training stage, the convergence performance of the model using P-Sigmoid is better than using Abs, because the gradients of P-Sigmoid are smoother. Furthermore, in the testing stage, the P-Sigmoid outperforms the Abs in the terms of MSE and MAE as shown in Figure 4(b). The preceding demonstrates that the performance gain of our model in terms of activation function is due to the proposed P-Sigmoid.

Ablation on density kernel fusion: We perform a series of experiments to determine the effect of density kernel size on the proposed density map generator. As the results shown in Figure 3, the small single kernel, e.g., $k = 5$, tends to yield better performance on the ShTech A and UCF-QNRF because their crowd scenarios are congested, while the large size, e.g., $k = 7$, can obtain better performance in relative sparse crowd scenarios, i.e., ShTech B. We thus choose to fuse 3×3 and 5×5 kernels in our density map generation framework, and achieve the best performance on all the datasets. This demonstrates that the suitable kernel fusion can improve the density map generation, because the fused density kernel has a better scene generalization.

6 Conclusion

In this paper, we primarily propose a novel vision transformer for modeling the humans’ Top-Down visual perception mechanism in the crowd counting task, as well as a density kernels fusion framework for obtaining more accurate ground-truth of density map from the dot annotation maps. Moreover, we also incorporate these two innovations into an adaptive crowd counting model, which can jointly learn density map estimator and generator within an end-to-end framework. Finally, we conduct extensive experiments, and prove our proposed approach can achieve superior performance in the terms of MAE and MSE on widely used datasets.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China(62106290) and the Program for Innovation Research in Central University of Finance and Economics.

References

[Alexey *et al.*, 2021] Dosovitskiy Alexey, Beyer Lucas, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Ashish *et al.*, 2017] Vaswani Ashish, Shazeer Noam, et al. Attention is all you need. In *Annual Conference on Neural Information Processing Systems*, pages 5998–6008, 2017.

[Boyu *et al.*, 2020] Wang Boyu, Liu Huidong, et al. Distribution matching for crowd counting. In *Annual Conference on Neural Information Processing Systems*, 2020.

[Chan *et al.*, 2008] Antoni B. Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[Chattopadhyay *et al.*, 2017] Prithvijit Chattopadhyay, Ramakrishna Vedantam, et al. Counting everyday objects in everyday scenes. In *Conference on Computer Vision and Pattern Recognition*, pages 4428–4437, 2017.

[Chen *et al.*, 2021] Binghui Chen, Zhaoyi Yan, et al. Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting. In *International Conference on Computer Vision*, pages 16065–16075, 2021.

[Dingkang *et al.*, 2021] Liang Dingkang, Chen Xiwu, et al. Transcrowd: Weakly-supervised crowd counting with transformer. *CoRR*, abs/2104.09116, 2021.

[Haroon *et al.*, 2018] Idrees Haroon, Tayyab Muhmmad, Athrey Kishan, et al. Composition loss for counting, density map estimation and localization in dense crowds. In *ECCV*, volume 11206, pages 544–559, 2018.

[Jia *et al.*, 2020] Wan Jia, Antoni B. Chan, et al. Modeling noisy annotations for crowd counting. In *Annual Conference on Neural Information Processing Systems*, 2020.

[Jia *et al.*, 2021] Wan Jia, Liu Ziquan, et al. A generalized loss function for crowd counting and localization. In *Conference on Computer Vision and Pattern Recognition*, pages 1974–1983, 2021.

[Lempitsky *et al.*, 2010] Victor S. Lempitsky, Andrew Zisserman, et al. Learning to count objects in images. In *Annual Conference on Neural Information Processing Systems*, pages 1324–1332, 2010.

[Lin *et al.*, 2021] Hui Lin, Xiaopeng Hong, et al. Direct measure matching for crowd counting. In *International Joint Conference on Artificial Intelligence*, pages 837–844, 2021.

[Liu *et al.*, 2021] Hao Liu, Qiang Zhao, et al. Bipartite matching for crowd counting with point supervision. In *International Joint Conference on Artificial Intelligence*, pages 860–866, 2021.

[Min *et al.*, 2008] Li Min, Zhang Zhaoxiang, Huang Kaiqi, and Tan Tieniu. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *International Conference on Pattern Recognition*, pages 1–4, 2008.

[Nicolas *et al.*, 2020] Carion Nicolas, Massa Francisco, et al. End-to-end object detection with transformers. In *ECCV*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229, 2020.

[Qi *et al.*, 2021] Wang Qi, Gao Junyu, et al. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *Trans. Pattern Anal. Mach. Intell.*, 43(6):2141–2149, 2021.

[Qingyu *et al.*, 2021] Song Qingyu, Wang Changan, et al. To choose or to fuse? scale selection for crowd counting. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 2576–2583, 2021.

[Shuai *et al.*, 2020] Bai Shuai, He Zhiqun, et al. Adaptive dilated network with self-correction supervision for counting. In *Conference on Computer Vision and Pattern Recognition*, pages 4593–4602, 2020.

[Song *et al.*, 2021] Qingyu Song, Changan Wang, et al. Rethinking counting and localization in crowds: A purely point-based framework. In *International Conference on Computer Vision*, pages 3365–3374, 2021.

[Tsung-Yi *et al.*, 2017] Lin Tsung-Yi, Dollár Piotr, et al. Feature pyramid networks for object detection. In *Conference on Computer Vision and Pattern Recognition*, pages 936–944, 2017.

[Wan and Chan, 2019] Jia Wan and Antoni B. Chan. Adaptive density map generation for crowd counting. In *International Conference on Computer Vision*, pages 1130–1139, 2019.

[Wan *et al.*, 2020] Jia Wan, Qingzhong Wang, and Antoni B Chan. Kernel-based density map generation for dense object counting. *Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[Yifan *et al.*, 2020] Yang Yifan, Li Guorong, et al. Weakly-supervised crowd counting learns from sorting rather than locations. In *ECCV*, pages 1–17, 2020.

[Yingying *et al.*, 2016] Zhang Yingying, Zhou Desen, et al. Single-image crowd counting via multi-column convolutional neural network. In *Conference on Computer Vision and Pattern Recognition*, pages 589–597, 2016.

[Yuhong *et al.*, 2018] Li Yuhong, Zhang Xiaofan, and Chen Deming. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018.

[Ze *et al.*, 2021] Liu Ze, Lin Yutong, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, pages 10012–1110022, 2021.

[Zheng *et al.*, 2021] Sixiao Zheng, Jiachen Lu, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.