Cross-Level Feature Relocation: Mitigating Information Loss in Cross-Layer Feature Fusion for Crowd Counting

Yuanyang Yin YYYIN@MAIL.USTC.EDU.CN and **Baoqun Yin** BQYIN@USTC.EDU.CN
University of Science and Technology of China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

In crowd counting, significant challenges persist due to scale variation, occlusion, and complex scene interference. Merging feature maps from different levels of the backbone network is an intuitive and efficient approach to addressing these issues. However, existing multi-scale merging algorithms often overlook a critical aspect: feature maps at different levels typically have varying resolutions, and traditional interpolation-based methods for feature fusion result in significant information loss, limiting the algorithm's multi-scale perception capability. To address this issue, we propose the Cross-Level Feature Relocation Module (CFRM), which regresses features across different levels into a unified representation space and utilizes a cross-level attention mechanism to transfer complementary information from low-resolution to high-resolution feature maps, significantly enhancing effective information utilization. Based on CFRM, we introduce the Cross-Level Feature Relocation Network (CFRNet), which exhibits strong multi-scale perception capabilities. Extensive experiments on five datasets and comprehensive ablation studies demonstrate the effectiveness of CFRM.

Keywords: Computer vision; Crowd counting; Mmulti-scale features

1. Introduction

Crowd counting is a significant branch of computer vision with valuable applications for its wide-ranging applications in public safety Li et al. (2013); Chaker et al. (2017); Onoro-Rubio and López-Sastre (2016); Kang et al. (2018), traffic monitoring Guerrero-Gómez-Olmedo et al. (2015), and agriculture Aich and Stavness (2017); Lu et al. (2017). The goal of crowd counting is to estimate the number or density of objects based on their features and distribution in images or videos. A major challenge in crowd counting is the significant scale variation and complex scene interference of the crowd in an image Song et al. (2021); Zhang et al. (2016), which complicates the problem since most datasets only provide point annotations.

In recent years, numerous researchers have proposed various solutions to address the challenges in crowd counting. Among these, a series of methods aimed at enhancing the receptive field have been developed to improve accuracy and robustness. These methods can be categorized as follows: 1) **Multi-column network** Zhang et al. (2016); Babu Sam et al. (2017) employ different kernel sizes or dilation rates for the convolution operations of various branches, allowing each branch to have a distinct receptive field and thereby enhancing the model's multi-scale perception capability. However, since the scale variation of counting objects in images is continuous and model parameters are limited, the number of branches is restricted. Consequently, this approach cannot fully adapt to the scale variation

problem. Additionally, there is significant redundancy among different branches, and the multi-column parallel structure makes the network architecture cumbersome. 2) **Multi-scale feature fusion methods** are currently the most widely used to obtain better multi-scale perception ability and have achieved remarkable resultsLin et al. (2017); Liu et al. (2018). However, this kind of fusion often ignores the correspondence between the features of each level and the scale of the counting objects, which leads us to only obtain a suboptimal solution. Moreover, the fusion process often requires upsampling the low-resolution features to achieve the alignment of the feature map sizes, which will damage the information in the hierarchical feature maps. 3) **Multi-level feature selection methods** achieve the task of crowd counting by identifying the most suitable features for objects of varying scales. This approach typically involves two branches: feature selection and result selection. In the study Song et al. (2021), density regression and confidence prediction are performed at each feature map level, with the optimal density regression selected adaptively based on confidence regression outcomes. This method effectively establishes a correspondence between feature map levels and object scales, yet it requires training an additional confidence branch to pinpoint the optimal density map patch for each regression level. Han et al. (2023) introduced STEERER, a method that utilizes features from various levels to perform crowd counting. STEERER acknowledges that feature maps at different levels exhibit varying sensitivity and perception abilities towards different objects. By selectively incorporating features from diverse levels, STEERER aims to derive an optimized feature map. However, whether in result or feature selection, aligning feature maps inevitably involves some degree of information degradation and loss in low-resolution feature maps. While this may not significantly impact feature maps with minor resolution discrepancies, it limits researchers from effectively integrating or selecting feature maps with substantial resolution differences to capture complementary information between upstream and downstream features.

In this study, we analyze how features at different levels demonstrate varied sensitivities in counting objects across different scales, indicating that each image patch corresponds to an optimal feature representation. Our objective is to select the most suitable feature representation for each counting object while preserving representation coherence across all feature maps, followed by performing crowd density regression. As illustrated in Figure 1, our approach involves transferring information from low-resolution feature maps to corresponding regions of high-resolution feature maps. We progressively refine feature selection through self-attention mechanisms to derive an optimal feature map for density map regression. To facilitate this process, we introduce a cross-level feature relocation module (CFRM). This module not only accurately aligns large-scale targets perceived in low-resolution feature maps with their high-resolution counterparts but also effectively utilizes rich semantic information from low-resolution features. Ensuring the effectiveness of the feature relocation process requires aligning features across all levels into a unified semantic space, achieved through the design of a regression head for each feature level. Additionally, we introduce an axial structure via CFRM to enhance upstream and downstream information fusion, thereby mitigating challenges associated with feature information loss during long-distance hierarchical feature decoding. Key contributions of this paper include:

- We propose an innovative method, the cross-level feature relocation module, to obtain optimal features for crowd counting. By relocating features from the low-resolution feature map to corresponding regions of the high-resolution feature map, we mitigate

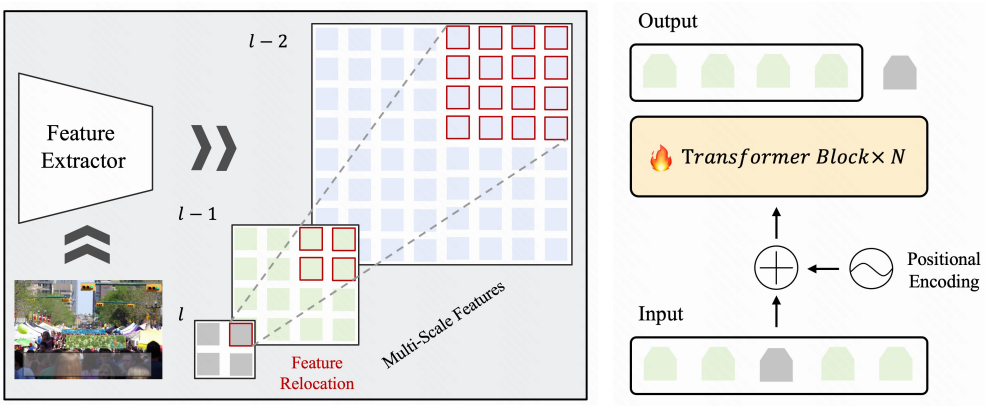

Figure 1: An illustration of the cross-level feature relocation. Due to the perspective effect, the head scale of the gray region in the input image is large, and the head scale of the green region is small. In layer l (gray), the features are more sensitive to the large-scale counting objects, while in layer $l-1$ (greed), the features are more sensitive to the small-scale counting objects. We first manually map the features of layer l to the feature map of layer $l-1$ by position encoding and then perform fine-grained adaptive feature relocation by window self-attentionLiu et al. (2021).

information loss during the decoding process and enhance our multi-scale perception capabilities. To our knowledge, this represents a significant advancement in using self-attention algorithms for multi-scale perception.

- We meticulously design CFRNet for the counting task, enabling the relocation of features from different levels into a unified feature map through CFRM. This approach allows us to derive tailored features for counting objects across various scales. Additionally, we employ CFRM to develop an axial decoding structure for CFRNet, effectively addressing issues related to information loss during hierarchical information decoding.

- We conducted extensive experiments on the popular crowd-counting datasets to verify the solid progress achieved by our method, including ShanghaiTech A, ShanghaiTech B JHU CROWD++, UCF_CC_50 and NWPU

2. Related Work

2.1. Multi-column based Method

Crowd counting methods based on multi-column convolutional neural networks aim to enhance multi-scale feature detection by employing multiple convolutional columns for feature extraction and regression. MCNN Zhang et al. (2016) utilizes three columns of networks

with varied convolutional kernel sizes to extract crowd features across different scales. These features are then aggregated to regress the crowd density map. Similarly, CrowdNet Boominathan et al. (2016) adopts a two-column convolutional network, integrating deep and shallow layers to capture high-level and low-level semantic information for predicting density maps from crowd images. In contrast to these approaches, which require multiple convolutional operations per input, SwitchCNN Babu Sam et al. (2017) introduces a classification branch that dynamically selects the most suitable convolutional column for each image. However, these methods do not fully address the challenge of scale variation inherent in crowd-counting tasks, where target objects exhibit continuous size variations. Moreover, the use of a multi-column parallel structure significantly increases both computational and storage overhead. This structure also introduces substantial information redundancy within each convolutional column.

2.2. Multi-Level Feature Fusion

Multi-level feature fusion methods capitalize on the diverse scale perception abilities embedded within various internal layers of convolutional neural networks. By integrating outputs from different network levels, these methods derive feature representations that are rich in content. Compared to multi-column approaches, they more effectively manage significant computational overhead, address shortcomings in single-column feature representations, and mitigate concerns of feature redundancy. These methods fall into two categories: deliberate fusion structures and attention mechanisms. SaCNN Zhang et al. (2018) extracts multi-level features from a single-column convolutional network and merges them to enhance multi-scale receptive abilities. MBTTBF Sindagi and Patel (2019) employs a Bottom-Top and Top-Bottom structure for hierarchical feature fusion, facilitating comprehensive semantic information from both ends. TeDNet Jiang et al. (2019) utilizes an encoder-decoder architecture with a densely connected cross-layer decoding structure, effectively combining features from different decoding stages for enhanced representation. In techniques employing attention mechanisms, AFN Zhang et al. (2019) and DSSINet Liu et al. (2019a) leverage conditional random fields to achieve hierarchical feature fusion, enabling precise integration of features across multiple levels. Hossain et al. Hossain et al. (2019) employ distinct perception modules to extract and integrate local and global scale information, integrating these with backbone network features to achieve adaptive multi-scale attention. These approaches collectively demonstrate advancements in leveraging multi-level feature fusion to enhance crowd-counting accuracy by effectively integrating diverse scale perception capabilities within convolutional neural networks.

2.3. Multi-Level Feature Selection

The idea of this type of method is that the density features of different levels are optimal for specific scale features, and global feature fusion can only produce suboptimal solutions. Therefore, this approach eschews the multi-scale fusion method to implement density regression via feature selection or density selection with the aim of achieving the optimal solution. For instance, Varior et al. Varior et al. (2019) adopt a novel soft attention mechanism operating at multiple scales to learn a set of gated masks and combine the density prediction outcomes generated by various levels into the ultimate result. Similarly, SASNet Song et al.

(2021) proposes a scale-adaptive selection network that automatically learns the correspondence between scales and feature levels. It trains the Confidence Branch by calculating the optimal solution corresponding to the patch to achieve the selection of different-level prediction results. Additionally, STEERER Han et al. (2023) proposes a method that focuses on selecting and inheriting hierarchical features instead of directly fusing prediction results at each level. This strategy prevents the loss of information caused by feature fusion. However, the challenge lies in choosing appropriate features and prediction outcomes at the patch level.

3. Our Approach

As shown in Figure 2, we first use a feature extraction network to obtain multi-level feature representations of the image with different resolutions, each with different sensitivity to counting objects of different scales. CFRNet employs CFRM (Cross-level Feature Reuse Module) to execute both hierarchical and axial decoding processes, culminating in a feature map that consolidates effective information across all levels with minimal information loss. We use distributed supervision to ensure that the feature representations at each level are in the same semantic space.

3.1. Multi-level Feature Representation

Convolutional neural networks possess a hierarchical structure, encompassing various levels that provide feature maps of different resolutions, corresponding to receptive fields of different scales, which are ideally suited to enhance the network's multi-scale perceptual capabilities. STEERER Han et al. (2023) demonstrates that lower-resolution feature maps are highly effective in capturing large objects, while higher-resolution feature maps detect small-scale objects. The task of crowd counting presents a major challenge due to the wide and continuous scale variation of the objects. Therefore, it is crucial to construct multi-level features with varying receptive fields. We use VGG16-BN Simonyan and Zisserman (2014) as our backbone network and supplement it with a global self-attention module (GAM). We extract the last three layers of output features from VGG16-BN to obtain hierarchical features with varying scale sensitivity. To derive the image's global perception feature, we connect our GAM to the last layer of VGG16-BN. Then, we use convolution operations to compress the channels of the various hierarchical feature maps and the global perception feature map to the same dimension. Specifically, the input image $I \in R^{3\times H\times W}$ and the hierarchical features $\left\{\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4\right\}$ are obtained by passing it through our proposed backbone neural network. The resolution of feature F_i is given by $(h_i, w_i) = (\frac{H}{2^{1+i}}, \frac{W}{2^{1+i}})$, and the channel number for each level is 256.

3.2. Cross-Level Feature Relocation

Scale feature fusion effectively enhances receptive fields to tackle scale variations and complex scene influences in crowd-counting tasks. However, directly interpolating, adding, or concatenating features is suboptimal, as it tends to degrade the original information of each feature map, especially when there's a significant resolution disparity, as illustrated in Figure 1. Previous studies have employed a progressive approach for selecting or fusing feature

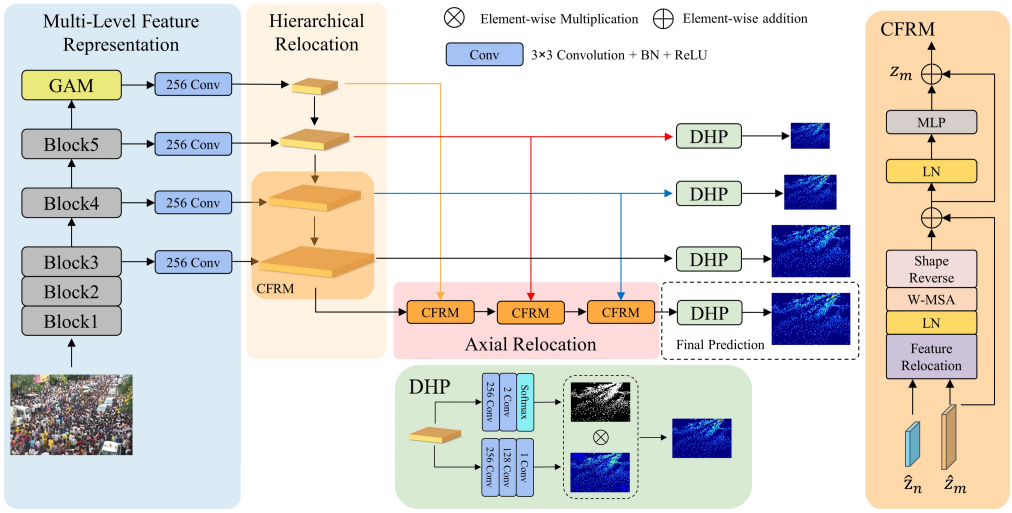

Figure 2: Overview of CFRNet. The proposed CFRNet consists of four components: multi-level feature extraction backbone, hierarchical decoder, axial decoder, and dual-head predictor. The input image is first processed through the backbone network to obtain feature representations at different resolutions. The hierarchical decoder is then used to enhance semantic information from top to bottom. The axial decoder is used to reduce information loss stemming from hierarchical decoding over long distances. A dual-headed predictor is implemented with a density regression branch and a foreground segmentation branch to get the density map.

maps. Nonetheless, this approach may lead to loss of upstream information, particularly when the number of features increases. To address the issue of feature information destruction and suboptimal solution problems caused by traditional cross-resolution feature fusion, we propose cross-level feature relocation. This approach can effectively solve these problems and reasonably achieve feature map fusion even when there are significant resolution differences. Cross-level feature relocation assumes that each feature vector in the low-resolution feature map perceives a larger field of view and extracts higher semantic information from the corresponding patch in the high-resolution feature map. Although the sensitivity to small-scale counting targets is decreased, the information in this scale feature map can perceive large-scale objects and also guide the low-resolution features from a higher-dimensional semantic information level. In counting tasks, we need both high-resolution feature maps to maintain the sensitivity and accurate localization of high-density small-scale counting objects and low-resolution feature maps with high-level semantic information and perception ability of large-scale counting objects to improve counting performance. To maintain feature integrity, CFRM avoids excessive manipulation of low-resolution features. Instead, each low-resolution feature vector is directly located on its corresponding patch in the high-resolution feature map. Subsequently, self-attention is employed to refine the localization

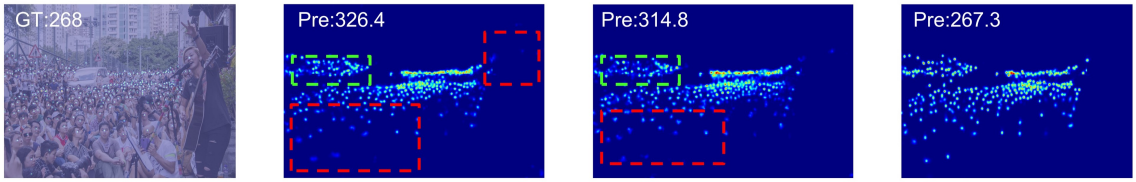

Figure 3: Feature fusion with different methods. Large counting objects come from low-resolution feature maps. Upsampling and then concatenating or summing will damage the feature information, resulting in low confidence in the final prediction results (red). For small-scale counting objects, features primarily originate from high-resolution feature maps. The implementation of a simplistic fusion method will result in interference from upstream invalid information (green). Our CFRNet effectively addresses these issues through the utilization of a feature relocation method.

of counting features and integrate high-level semantic information. Considering the higher computational cost of the global self-attention mechanism and the fact that the feature relocation process does not require global information, we employ the window self-attention approach for cross-level feature fusion.

Specifically, the calculation process for relocating features in CRFM is as follows:

$$Q = MLP\left(LN\left(C\left[\hat{\mathbf{z}}_m, \hat{\mathbf{z}}_n\right]\right)\right) \tag{1}$$

$$K = MLP\left(LN\left(C\left[\hat{\mathbf{z}}_m, \hat{\mathbf{z}}_n\right]\right)\right) \tag{2}$$

$$V = MLP\left(LN\left(C\left[\hat{\mathbf{z}}_m, \hat{\mathbf{z}}_n\right]\right)\right)\left[:, 0:w^2, :\right] \tag{3}$$

$$A = \left(softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) + C\left[B_m, B_n\right]\right)[:, 0:w2, :] \tag{4}$$

$$z_m = f\left(AV + \hat{\mathbf{z}}_m\right) \tag{5}$$

The C denotes the feature concatenation operation, $\hat{\mathbf{z}}_m$ and $\hat{\mathbf{z}}_n$ represent the tokens within the corresponding windows of F_m and F_n, w denotes the window size in the cross-level feature relocation process, B_m denotes the relative position embedding of $\hat{\mathbf{z}}_m$, and B_n denotes the position information of $\hat{\mathbf{z}}_n$ in $\hat{\mathbf{z}}_m$. More specifically, for the feature maps F_m and F_n, let their corresponding resolutions be $\left(\frac{H}{2^{1+m}}, \frac{W}{2^{1+m}}\right)$ and $\left(\frac{H}{2^{1+n}}, \frac{W}{2^{1+n}}\right)$. Then, for the tokens $\hat{\mathbf{z}}_m$ in a window of size (w, w) of F_m. According to Liu et al. (2021), we can obtain the relative position embedding for each position as B_m. The corresponding window size in the low-resolution feature map is $\left(\frac{w}{2^{n-m}}, \frac{w}{2^{n-m}}\right)$, and the tokens are $\hat{\mathbf{z}}_n$. We obtain the preliminary repositioning information B_n by calculating B_m using the nearest interpolation and then achieve the precise positioning of the feature information by self-attention.

Additionally, the CFRM module is lightweight, introducing only $12 \times c^2$ additional parameters, where c is the feature dimension. In terms of computational complexity, for

two feature maps of sizes (h, w, c) and $(h/2^i, w/2^i, c)$ with a relocation window size of (k, k), the additional computational cost of a single CFRM module can be expressed as follows:

$$\text{Compute} = \frac{h}{k} \times \frac{w}{k} \times 24 \times m \times c^2 + 4m^2c \tag{6}$$

$$m = k^2 + \frac{k^2}{4^i} \tag{7}$$

The total computational cost depends on the resolution of the input CFRM feature maps, the feature dimension, and the size of the window partitioning.

3.3. Hierarchical and Axial Feature Relocation

The input image undergoes processing through the backbone network to extract multi-level feature representations with varying perceptual capabilities for detecting objects of different scales. For CFRNet, we have designed a progressive decoding structure and an axial decoding structure to derive the feature map necessary for density regression. While the hierarchical nature of convolutional neural networks in the backbone network provides feature representations at different resolutions, it also results in shallow features that may not sufficiently preserve feature validity due to limited network depth. To address this, we propose initially employing the feature pyramid structure Lin et al. (2017) to conduct a top-down progressive fusion of feature information. This approach leverages high-level semantic guidance to iteratively enhance semantic information downstream. Mathematically, this process can be described as follows:

$$\begin{cases} F_i = \hat{F}_i & i = 4 \\ F_i = f\left(\text{CFRM}\left(\hat{F}_i, F_{i+1}\right)\right) & i = 1, 2, 3 \end{cases} \tag{8}$$

Where f represents the convolution operation, $\hat{F}_i, i = 1, 2, 3$ represent the hierarchical feature representations acquired from the backbone network, and $\hat{F}_4$ represents the global perception attention feature map obtained via the GAM. CFRM refers to the cross-level feature relocation process. In the process of progressively merging features, as the depth increases, upstream information will encounter a greater forgetting problem. This leads to a range of issues, including significant loss of feature information for counting objects and inadequate use of high-level semantic information. Therefore, we utilize CFRM to construct an axial decoding structure, guaranteeing full and effective utilization of information on every level. For the F_1 that directly generates the density map, we utilize the scale attributes of each level to enhance the information of this feature map. Involving: 1) Relocating the large-scale features to the high-resolution feature map via CFRM to enhance multi-scale perception ability. 2) Utilizing high-level semantic information to further augment the information in the feature map, thus addressing the issue of information loss attributed to the progressive feature fusion process. The computational process can be described as follows:

$$\begin{aligned} F_5 &= f\left(CFRM\left(F_1, F_4\right)\right) \\ F_5 &= f\left(CFRM\left(F_5, F_3\right)\right) \\ F_5 &= f\left(CFRM\left(F_5, F_2\right)\right) \end{aligned} \tag{9}$$

Where f represents the convolution operation, $F_i, i = 1, 2, 3, 4$ represents the features acquired through the Eq. (8), while the feature map of axial cross-level feature repositioning is obtained by executing Eq. (9).

3.4. Dual-Head Predictor

To minimize the influence of background information on density map regression predictions, we consulted Modolo et al. (2021) and devised a DHP module to enable crowd density map regression. The DHP module trains the task of segmenting the foreground and regressing density concurrently via a multi-task learning method. Additionally, it bolsters the model's capacity to withstand background interference by utilizing the foreground segmentation outcomes on the regressed density map. Caruana (1997) also suggests that learning shared representations between tasks and exploring particular region information from the training signal, can enhance the generalization performance of individual tasks.

3.5. Distributed supervision

The feature maps at varying levels show distinct sensitivity in detecting objects of different scales. To reposition high-level semantic information from the low-resolution feature map to the high-resolution feature map, we apply CFRM to develop a unified method for progressive decoding and axial decoding. The special positioning premise requires that the detections exist within the same semantic space. In other words, when a person is detected in that position, the feature map vector must have identical or comparable feature representation. To satisfy this requirement, we employ a distributed supervision technique to create a distinct DHP for each level of the feature map, executing density regression to guarantee that the semantic information of each level of the feature map is within the same semantic space. Specifically, we utilize DHP for density map regression and foreground segmentation map regression on $\{F_1, F_2, F_3, F_5\}$ separately, yielding four density regression results. It is noteworthy that a regression head was not intentionally designed for feature F_4, as this feature map is attained through the global self-attention operation, and its purpose is to furnish the network with global perceptual capability. Therefore, we do not enact regression supervision on it.

3.6. Loss Function

By the distributed loss in front of the DHM module, we separately monitor the foreground segmentation output and the regressed density map of the four DHP. To measure the loss for the density map regression branch, we employ the Euclidean distance between the predicted and ground truth density maps. For the foreground and background segmentation branch, we calculate the loss through cross-entropy between the foreground segmentation image and ground truth. Let X_j be the j-th image in a minibatch, $\hat{D}_i\left(X_j;\Theta\right)$ denotes the density map predicted by the i-th prediction head, D_j^i denotes the ground truth density map for the i-th prediction head. Let Θ represent the trainable parameters in the model, while $\hat{S}_i\left(X_j;\Theta\right)$ represents the foreground segmentation map predicted by the i-th prediction head and S_j^i denotes the ground truth segmentation map for the i-th prediction head. This signifies that the density loss function L_{mse}^i and the foreground segmentation loss function L_{ce}^i for the

i-th prediction head may be expressed as:

$$L_{ce}^{i}\left(\Theta\right)=-\frac{1}{N}\sum_{j=1}^{N}y_{j}^{i}log\left(\hat{S}_{i}\left(X_{j};\Theta\right)\right) \tag{10}$$

$$L_{mse}^{i}\left(\Theta\right)=\frac{1}{N}\sum_{j=1}^{N}||\hat{D}_{i}\left(X_{j};\Theta\right)-D_{j}^{i}|\,|_{2}^{2} \tag{11}$$

Where N represents the number of samples involved in the training process, y_j^i denotes the value assigned to each pixel in the ground truth foreground segmentation map S_j^i. The final loss is obtained by summing the corresponding density regression loss and segmentation loss for each prediction head. The final loss function is calculated by:

$$L=\sum_{i\in\{1,2,3,5\}}\left(L_{mse}^{i}\left(\Theta\right)+\gamma L_{ce}^{i}\left(\Theta\right)\right) \tag{12}$$

Where γ is a ratio factor, used to balance the training of the two branches. After validation, the experiment process of this paper sets $\gamma = 1$.

4. Experiments

To showcase the efficacy of CFRNet, we conducted experiments on four demanding datasets: SHHA, SHHB, JHU-CROWD++, UCF_CC_50, and NWPU. Additionally, we conducted detailed ablation studies to demonstrate the effectiveness of each module. We generated ground-truth density maps using a Gaussian kernel with a fixed sigma of 4. Mean Absolute Error (MAE) and Mean Squared Error (MSE) served as evaluation metrics.

4.1. Experimental details

The backbone network of our encoder comprises the first 13 convolutional layers of VGG-16_bn Simonyan and Zisserman (2014), which is pre-trained on ImageNet. For multi-level feature extraction, we utilize the pre-trained VGG-16_bn first 13 convolutional layers in conjunction with our GAM. In the hierarchical decoding process, we use a window size of 8 for feature relocalization. In the process of axial decoding, we sequentially relocate the features of $\{F_4, F_3, F_2\}$ to F_1 using window sizes of 16, 8 and 8 respectively. To train the model, we randomly crop four image patches of 256×256 from four images and horizontally flip the images with a 0.5 probability while grayscale them with a probability of 0.1. We optimize the model with Adam, setting the learning rate to 1e-5.

4.2. Ablation Study

We conduct ablation experiments on SHHA to validate the effectiveness of each module and structure.

Effectiveness of the CFRM. In Table 1, we first remove CFRM and replace the feature relocation process with alignment addition or concatenation operations. We obtain MAE of 54.6 and 54.1, while the method using CFRM can achieve an MAE of 51.1. The

Table 1: Comparing various approaches to feature fusion. This study employs a structure that combines hierarchical decoding and axial decoding strategies.

Method	MAE	MSE
CFRNet + Add	54.6	90.3
CFRNet + Concat	54.1	88.2
CFRNet + CFRM	51.1	85.2

error rate is reduced by 6.4% and 5.5% compared to the former. Removing the feature relocation function of CFRNet, and the destruction and loss of high-level semantic information will be severely aggravated, and the feature relationship between the hierarchical features and the counting object scale cannot be effectively established.

Table 2: Ablation experiments for each module in CFRNet on SHHA.We remove the GAM and DHP modules from CFRNet as the baseline network for comparison.

Method	MAE	MSE
CFRNet w/o GAM & DHP	58.6	96.4
CFRNet w/o GAM	53.5	89.5
CFRNet w/o DHP	55.1	90.1
CFRNet	51.1	85.2

Effectiveness of the GAM. The role of the GAM is to give CFRNet the ability to perceive globally and distinguish between the background and the crowd, thereby improving the network's anti-interference ability. Removing GAM severely damages CFRNet's overall perception, leading to incorrect identification of backgrounds as crowd areas. In Table 2, when adding GAM to the baseline network, the MAE indicator increases by 6.0% to reach 55.1, while the MSE indicator increases by 6.5% to reach 90.1.

Effectiveness of the DHP. DHP divides the density regression task into two parts: foreground segmentation and density regression. Thanks to the advantages of distributed supervision, the dual-head prediction method can improve the counting performance in a simple and effective way, and make each branch more focused on the current sub-task. When we add the DHP module to the baseline network, the MAE indicator reaches 53.5, an increase of 8.7%, and the MSE indicator reaches 89.5, an increase of 7.2%.

Effectibeness of the Hierarchical and Axial Decoder. We believe that the downstream features with greater resolution are incapable of extracting higher-level semantic information due to insufficient network depth, so we design a hierarchical progressive decoding structure to enhance the downstream information from top to bottom. However, the long-distance progressive decoding process results in the loss of upstream information, preventing low-level information from being guided by upstream high-level semantic information. Therefore, the axial decoding structure designed by CFRM is utilized to attain cross-level decoding, which can decode cross-level features directly without damaging the

Table 3: Comparing the effectiveness of hierarchical decoder and axial decoder. Where DH stands for hierarchical decoder, and AD represents axial decoder.

Method	MAE	MSE
CFRNet w/o HD&AD	68.7	117.0
CFRNet w/o HD	55.6	91.1
CFRNet w/o AD	54.8	89.2
CFRNet	51.1	85.2

Table 4: Performance comparison with state-of-the-art methods on SHHA, SHHB, JHU-Crowd++, UCF_CC_50, and NWPU

Methods	SHHA		SHHB		JHU-Crowd++		UCF_CC_50		NWPU	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN Zhang et al. (2016)	110.2	173.2	26.4	41.3	188.9	483.4	377.6	509.1	232.5	714.6
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
CAN Liu et al. (2019b)	62.3	100.0	7.8	12.2	100.1	314.0	212.2	243.7	106.3	368.5
S-DCNet Xiong et al. (2019)	58.3	95.0	6.7	10.7	-	-	204.2	301.3	-	-
BL Ma et al. (2019)	62.8	101.8	7.7	12.7	75.0	299.9	229.3	308.2	-	-
ADSCNet Bai et al. (2020)	55.4	97.7	6.4	11.3	-	-	198.4	267.3	-	-
ASNet Jiang et al. (2020)	57.8	90.1	-	-	-	-	174.8	251.6	-	-
CLTR Liang et al. (2022)	56.9	95.2	6.5	10.6	59.5	240.6	-	-	74.3	333.8
ChfL Shu et al. (2022)	57.5	94.3	6.9	11.0	57.0	**235.7**	-	-	-	-
CLH Wang et al. (2022)	59.4	93.5	7.4	11.4	62.4	248.9	-	-	86.7	435.0
GauNet Cheng et al. (2022)	54.8	89.1	6.2	9.9	58.2	245.1	186.3	256.5	-	-
STEERER Han et al. (2023)	54.5	86.9	**5.8**	**8.5**	**54.3**	238.3	-	-	**63.7**	**309.8**
MDKNet Guo et al. (2024)	55.4	91.6	6.4	10.0	-	-	-	-	66.7	314.0
Gramformer Lin et al. (2024)	54.7	87.1	-	-	53.1	228.1	-	-	72.5	316.4
CLIP-EBC (ResNet50) Ma et al. (2024)	55.0	88.7	6.3	10.2	-	-	-	-	-	-
Ours	**51.1**	**85.2**	6.4	10.9	58.5	252.4	**174.1**	**237.0**	67.5	312.4

upstream information characteristics. Table 3 presents ablation experiments for both the hierarchical decoding structure and axial decoding structure. We utilize CFRNet as the base network without its hierarchical decoder and axial decoder. It is observable that the MAE indicator can reach 54.8 when the network is designed using only the hierarchical decoder. Similarly, the MAE indicator can reach 55.6 when the axial decoder is carried out. Utilizing both the hierarchical decoding structure to enhance downstream features and the axial decoding structure to avoid neglecting upstream information results in an MAE of 51.1 and an MSE of 85.2, which represents a 25.6% and 27.2% increase over the baseline network, respectively.

4.3. Comparisons with State-of-the-Arts

To demonstrate the effectiveness of CFRNet, we compare its performance with state-of-the-art methods on four challenging datasets. The results are presented in Table 4. The best performance is indicated by bold numbers, and the second-best performance is indicated by underlined numbers.

ShanghaiTech Dataset. The ShanghaiTech Dataset Zhang et al. (2016) has two parts: Part-A and Part-B. Part-A contains 300 training and 182 testing images, with higher density and crowding, while Part-B includes 400 training and 316 testing images from busy streets, featuring sparser scenes. CFRNet achieves top performance on Part-A and second-best on Part-B, especially in the congested scenarios of Part-A.

JHU-Crowd++. The JHU-Crowd++ dataset Sindagi et al. (2020) is a large-scale collection featuring 4,372 images and 1,515,005 head annotations, divided into 2,772 for training, 500 for validation, and 1,600 for testing. To standardize the dataset, we limit the image resolution to 1024, which may lead to some information loss; however, we still achieve a notable performance with a mean absolute error (MAE) of 58.5.

UCF_CC_50. The UCF_CC_50 Idrees et al. (2013) is a small dataset for crowd counting, with 50 images of crowds from different websites covering different scenarios and environmental conditions. The crowd density in the dataset is very high, with an average of 1279 heads per image and a maximum of 4633 heads. We experimented with CFRNet using a five-fold cross-validation approach and achieved the best performance, where the MAE reached 174.1 and the MSE reached 237.0.

NWPU. The NWPU dataset Wang et al. (2020) is a widely used benchmark for crowd counting research, featuring diverse scenes captured by surveillance cameras across various real-world scenarios, including streets, parks, and public events. With precise head count annotations, NWPU is essential for evaluating algorithms in crowd density estimation and counting. Its comprehensive and challenging nature makes it a valuable resource for advancing crowd counting methodologies. Our results, with a mean absolute error (MAE) of 67.5 and a mean squared error (MSE) of 312.4, demonstrate competitive performance in the field.

5. Conclusion

In this paper, we propose CFRNet, which effectively mitigates information loss from fusing features at different resolutions in existing crowd counting algorithms when addressing scale variation, through a feature relocation algorithm. CFRNet uses CFRM to create a hierarchical decoding structure that selects optimal features for density regression at each scale while preserving multi-scale representation. The axial decoding structure, implemented via CFRM, addresses information loss from stage-wise decoding. We recommend distributed supervision to ensure features at each level remain in the same semantic space during relocation. Comprehensive tests on four challenging datasets demonstrate the significance of our approach.

References

Shubhra Aich and Ian Stavness. Leaf counting with deep convolutional and deconvolutional networks. In *ICCV*, 2017.

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

Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4594–4603, 2020.

Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 640–644, 2016.

Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

Rima Chaker, Zaher Al Aghbari, and Imran N Junejo. Social network model for crowd anomaly detection and localization. *Pattern Recognition*, 61:266–281, 2017.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19638–19648, 2022.

Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. Extremely overlapping vehicle counting. In *IbPRIA*, 2015.

Mingyue Guo, Binghui Chen, Zhaoyi Yan, Yaowei Wang, and Qixiang Ye. Virtual classification: Modulating domain-specific knowledge for multidomain crowd counting. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Tao Han, Lei Bai, Lingbo Liu, and Wanli Ouyang. Steerer: Resolving scale variations for counting and localization via selective inheritance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21848–21859, 2023.

Mohammad Hossain, Mehrdad Hosseinzadeh, Omit Chanda, and Yang Wang. Crowd counting using scale-aware attention networks. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1280–1288. IEEE, 2019.

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

Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4706–4715, 2020.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6133–6142, 2019.

Di Kang, Zheng Ma, and Antoni B Chan. Beyond counting: Comparisons of density maps for crowd analysis tasks—counting, detection, and tracking. *TCSVT*, 2018.

Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36 (1):18–32, 2013.

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

Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pages 38–54. Springer, 2022.

Hui Lin, Zhiheng Ma, Xiaopeng Hong, Qinnan Shangguan, and Deyu Meng. Gramformer: Learning crowd counting via graph-modulated transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3395–3403, 2024.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1774–1783, 2019a.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

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

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

Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. Tasselnet: counting maize tassels in the wild via local counts regression network. *Plant methods*, 13(1):1–17, 2017.

Yiming Ma, Victor Sanchez, and Tanaya Guha. Clip-ebc: Clip can count accurately through enhanced blockwise classification. *arXiv preprint arXiv:2403.09281*, 2024.

Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6142–6151, 2019.

Davide Modolo, Bing Shuai, Rahul Rama Varior, and Joseph Tighe. Understanding the impact of mistakes on background regions in crowd counting. In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1650–1659, 2021.

Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *ECCV*, 2016.

Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19618–19627, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Vishwanath A Sindagi and Vishal M Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1002–1012, 2019.

Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2594–2609, 2020.

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, pages 2576–2583, 2021.

Rahul Rama Varior, Bing Shuai, Joseph Tighe, and Davide Modolo. Multi-scale attention network for crowd counting. *arXiv preprint arXiv:1901.06026*, 2019.

Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *TPAMI*, 2020.

Qi Wang, Juncheng Wang, Junyu Gao, Yuan Yuan, and Xuelong Li. Counting like human: Anthropoid crowd counting on modeling the similarity of objects. *arXiv preprint arXiv:2212.02248*, 2022.

Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8362–8371, 2019.

Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Attentional neural fields for crowd counting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5714–5723, 2019.

Lu Zhang, Miaojing Shi, and Qiaobo Chen. Crowd counting via scale-adaptive convolutional neural network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1113–1121. IEEE, 2018.

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