Heterogeneous Dual-Attentional Network for WiFi and Video-Fused Multi-Modal Crowd Counting

Lifei Hao, Baoqi Huang, *Senior Member, IEEE*, Bing Jia, *Member, IEEE*, and Guoqiang Mao, *Fellow, IEEE*

***Abstract*—Crowd counting aims to estimate the number of individuals in targeted areas. However, mainstream vision-based methods suffer from limited coverage and difficulty in multi-camera collaboration, which limits their scalability, whereas emerging WiFi-based methods can only obtain coarse results due to signal randomness. To overcome the inherent limitations of unimodal approaches and effectively exploit the advantage of multi-modal approaches, this paper presents an innovative WiFi and video-fused multi-modal paradigm by leveraging a heterogeneous dual-attentional network, which jointly models the intra- and inter-modality relationships of global WiFi measurements and local videos to achieve accurate and stable large-scale crowd counting. First, a flexible hybrid sensing network is constructed to capture synchronized multi-modal measurements characterizing the same crowd at different scales and perspectives; second, differential preprocessing, heterogeneous feature extractors, and self-attention mechanisms are sequentially utilized to extract and optimize modality-independent and crowd-related features; third, the cross-attention mechanism is employed to deeply fuse and generalize the matching relationships of two modalities. Extensive real-world experiments demonstrate that our method can significantly reduce the error by 26.2%, improve the stability by 48.43%, and achieve the accuracy of about 88% in large-scale crowd counting when including the videos from two cameras, compared to the best WiFi unimodal baseline.**

***Index Terms*—Crowd counting, passive WiFi sensing, visual information, multi-modal fusion, deep learning.**

I. Introduction

As CITIES rapidly expand, monitoring gathering crowds has become an important research field. Despite advancements, accurately estimating crowd counts in large-scale surveillance areas remains a challenging open problem, essential for numerous applications [1], [2], [3] including crowd management, traffic control, urban planning, and public safety. Recent tragic incidents, such as the stampedes in Itaewon, South Korea [4], and Kanjuruhan Stadium, Indonesia [5], highlight the urgent need for precise crowd monitoring to preemptively identify emergencies and implement effective safety measures.

To accurately count crowds in a scenario, the most straightforward approach involves utilizing visual information. Consequently, contemporary studies primarily extract pedestrian features from images or video frames through detection or regression techniques [6]. However, this approach is significantly impacted by variable lighting conditions, occlusions, and scale changes, rendering it unreliable for robust applications [7]. More critically, a single camera cannot encompass extensive surveillance areas, severely limiting the applicability. As such, algorithms that analyze radio frequency (RF) signals have been developed, which estimate crowd sizes by exploiting the correlation between wireless signals and pedestrian numbers [8]. Among these, the WiFi Channel State Information (CSI)-based method is notable. It precisely correlates detailed WiFi signal attributes with crowd counts and, with the aid of deep learning, achieves accuracies exceeding 85% [9], [10]. Despite its effectiveness, this method's practicality is restricted to smaller, controlled indoor environments.

To break through scalability limitations, emerging research on passive WiFi sensing-based crowd counting deploys a special kind of access point (AP), termed WiFi sniffer, to passively sense the existence of pedestrians in the scenario by capturing and parsing the probe (request) frames sent from their mobile devices [11]. However, the challenges from persons with multiple WiFi-enabled mobile devices, uncertain sniffing, and MAC address randomization [12] severely affect this kind of method and make it only get a rough accuracy of less than 80%, which is not sufficient to support high-precision crowd counting in large-scale surveillance areas.

Multi-modal deep learning (MMDL) is a highly active interdisciplinary field focused on developing models capable of processing and correlating diverse information types, thereby delivering more accurate estimates or predictions than traditional unimodal approaches [13]. Inspired by recent advancements in MMDL and the complementarity between WiFi and video modalities in terms of scalability, accuracy, and both deployment and computational costs, we propose an innovative approach to leverage a network of pervasive WiFi sniffers and

Manuscript received 5 November 2023; revised 21 June 2024; accepted 31 July 2024. Date of publication 15 August 2024; date of current version 5 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62262046, Grant 42161070, and Grant U21A20446, in part by the Natural Science Foundation of Inner Mongolia A.R. of China Grant 2021ZD13 and Grant 2023MS06004, in part by the Science & Technology Plan Project of Inner Mongolia A. R. of China under Grant 2022YFSJ0027, Grant 2021GG0163, and Grant 2023KJHZ0016, in part by the Ordos Science & Technology Plan Grant YF20240029, in part by the University Youth Science and Technology Talent Development Project (Innovation Group Development Plan) of Inner Mongolia A. R. of China under Grant NMGIRT2318, in part by the fund of supporting the reform and development of local universities (Disciplinary construction), and in part by the fund of First-class Discipline Special Research Project of Inner Mongolia A. R. of China under Grant YLXKZX-ND-036. Recommended for acceptance by A. Conti. *(Corresponding author: Baoqi Huang.)*

Lifei Hao, Baoqi Huang, and Bing Jia are with the Engineering Research Center of Ecological Big Data, Ministry of Education, the Inner Mongolia A.R. Key Laboratory of Wireless Networking and Mobile Computing, and the College of Computer Science, Inner Mongolia University, Hohhot 010021, China (e-mail: cshbq@imu.edu.cn).

Guoqiang Mao is with the Research Institute of Smart Transportation, Xidian University, Xi'an 710071, China.

Digital Object Identifier 10.1109/TMC.2024.3444469

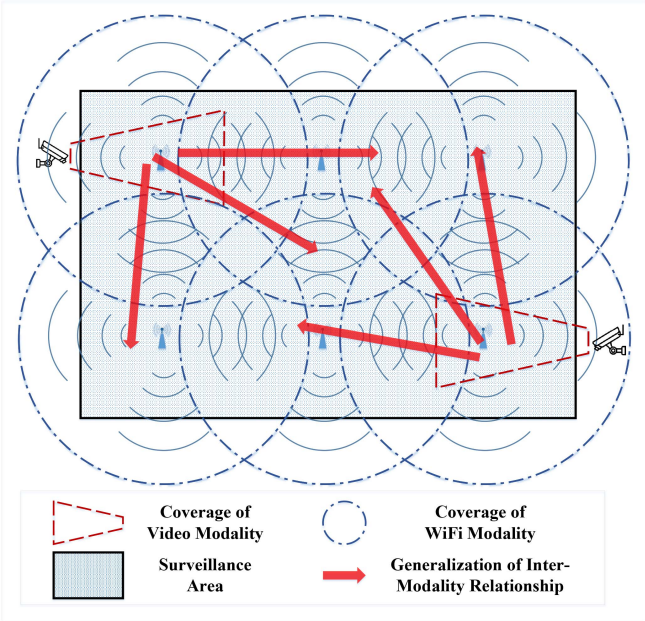

Fig. 1. Illustration of the deployment of multi-modal HSN.

select video cameras to capture distinct sets of multi-modal data, and an MMDL-based technique to intricately merge these diverse data streams, enabling the adaptive delivery of either global or local crowd counts, which can addresses the limitations inherent in unimodal solutions and bridges the gap in achieving high-precision, multi-modal crowd counting.

Accordingly, this paper introduces a flexible, easily deployable Hybrid Sensing Network (HSN), depicted in Fig. 1, to capture multi-modal data via passive WiFi sensing and camera systems. Additionally, we develop a Heterogeneous Dual-Attentional Network (HDANet) within a unified MMDL framework to model both intra- and inter-modality relationships. First, the synchronously collected multi-modal data is differentially preprocessed to eliminate noise and maximally reserve the crowd-related information, and then fed into a convolutional neural network (CNN)-based feature extraction modules with heterogeneous architectures to extract different levels and perspectives features represented by each modality. Second, inspired by the existing MMDL models for other tasks [14], [15], we design two effective attention modules, namely the self-attention and cross-attention modules, to model the intra- and inter-modality relationships, respectively. In the former, features are independently embedded with positional information and fed into the Transformer encoder [16], to discover intra-modality relationships and achieve context aggregation. In the latter, we employ the features of global WiFi modality as the key and value, and the features of multiple local video modalities as the query, then utilize scaled dot-product attention to achieve local matching of inter-modality relationships and the feed-forward neural network (FFN) to generalize these relationships through the whole feature map. Third, the multi-layer perceptron (MLP) is employed to fuse the concatenated self-attention and cross-attention features, and to adaptively select the best features for local and global crowd counting.

To evaluate our proposed multi-modal crowd counting method, an experimental HSN is deployed in a real campus environment with the surveillance area of about 4000 m^2, and a multi-modal dataset is collected during peak hours after classes. Extensive experimental results demonstrate our method can significantly reduce the counting error, improve the counting stability and achieve the state-of-the-art accuracy. Furthermore, multiple ablation studies not only confirm the effectiveness of our unimodality treatments and each component of the HDANet, but also extensively investigate the diversities of different fusion modes and the impact of multiple aspects, offering valuable guidance for the practical application and optimization of our method.

In summary, our main contributions are three-fold:

- *Innovative Multi-Modal Paradigm:* We introduce and realize all respects of fusing significantly heterogeneous modalities (i.e., WiFi and Video) for large-scale crowd counting from scratch, which exploits the complementary features between two modalities;
- *HDANet Fusion Model:* We develop a novel MMDL-based fusion model, termed HDANet, which differentially preprocessing sensing data and extract features from disparate modalities, and then meticulously fuses both intra- and inter-modality correlations;
- *Real-World Application and Insights:* We extensively validate the feasibility and effectiveness of both the paradigm and model in a real-world, large-scale scenario, offering practical insights and analysis based on our findings.

The rest of this paper is organized as follows. Section II surveys the related literature. Section III completely elaborates the proposed method. Section IV presents extensive experimental results and analyses. Section V concludes the whole paper and sheds light on the future work.

II. Related Work

In this section, we shall briefly introduce the literature of crowd counting and fusion of WiFi and videos.

A. Unimodality-Based Crowd Counting Method

The unimodal approach to crowd counting typically encompasses vision-based and WiFi-based methods. Early vision-based methods [17] extract pixel-level or texture-level shallow features to identify individuals, resulting in rough results; individual recognition [18] can achieve accurate results but are only suitable for sparse scenarios; line counting [19] counts pedestrians crossing the marked line and cannot identify stopping ones; recent density mapping [20], [21], [22], [23] suffers from the scale change and video quality [24]. Overall, vision-based methods though are accurate, but are limited by the inherent limitations of visual modality [25]. In contrast, the low-cost and strong-scalability passive WiFi sensing (PWS) enables large-scale crowd counting [26]. Fukuzaki et al. [27] verified the feasibility of PWS-based crowd counting through field experiments, with an error rate of over 30%; similarly, Weppner et al. [28] employed WiFi sniffers with directional antennas to reduce the rate to nearly 20%; our recent work [29]

demonstrated the effectiveness of deep learning in capturing the complex spatial-temporal relationship between WiFi sensing data and crowd counts. However, the existing PWS-based methods, although seem feasible in practice, are limited by relatively low accuracy.

In summary, different from existing unimodal approaches, we proposed to utilize the complementarities between the two modalities, especially the scalability and accuracy, to deeply mine and fuse the crowd-related information in a multi-modal paradigm.

B. Multi-Modality-Based Crowd Counting

Though the current MMDL studies mainly focus on the fusion of natural language [30], vision [31] and audio [32], a few recent studies attempted to shift MMDL to crowd counting. First, researchers conducted the crowd counting study based on the fusion of images and image-like modalities. Wagner et al. [33] leveraged two CNNs to extract features from RGB and thermal images, and then fuse them by fully connected layers, reducing the impact of lighting conditions; Schlosser et al. [34] utilized a similar approach to fuse RGB and LiDAR modalities, and reported a superior result compared to pure RGB images; Zhang et al. [35] proposed a plug-and-play cross-modality spatial-channel attention module for RGB-Thermal or RGB-Depth fusion. Such methods still suffer from the shallow fusion of images of different forms and the inherent limitations of the vision-based approach. Second, some researchers tried to leverage MMDL based on vision unimodality. Hydra-CNN [36] estimates overall crowd density by learning a multi-scale nonlinear regression model for scenarios with any scale; MoCNN [37] integrates multiple CNN models, which can adaptively select appropriate ones for processing different appearances of scenarios, and then weighs and sums the respective estimates. However, such methods do not fuse new modalities and strictly do not belong to the multi-modal fusion.

In summary, there has yet to be any work on fusing sensing data from significantly heterogeneous modalities, and our systematic research on the WiFi and video-fused paradigm can effectively dig the enormous potential of MMDL for large-scale and high-precision crowd counting.

C. Fusion of WiFi and Video Modalities

The complementarity between WiFi and video modalities in terms of location estimation has attracted the attention of researchers in recent years [38], [39]. In pedestrian tracking, an early study [40] proposed a switching-based mechanism to fuse measurements obtained from WiFi sniffers and a camera for tracking single pedestrians; MOLTIE system [41] further employed trajectory correlation-based matching approach to extend in multi-pedestrian case. In indoor localization, WAIPO system [42] utilized various build-in sensors of smartphones, especially the camera and WiFi, and employed a probability-based weighting method to fuse WiFi localization and image matching results to obtain location estimate; Redžić et al. [43] designed two result-level fusion mechanisms based on threshold and particle filter, respectively, to fuse similar WiFi and image localization results; on these grounds, Tang et al. [44] further obtained fused localization based on Bayesian probability weighting of two estimates, and then leveraged the hybrid whale optimization algorithm (HWOA) to adaptively determine the threshold for fusing three estimates. Besides the field restriction, the existing fusions are essentially based on late fusion, and thus cannot deeply mine the relationship between two modalities.

In summary, our work not only extends the active localization or tracking into the field of passive crowd counting, but also proposes a more effective middle fusion-based mechanism with dual-attention embedding to discover and exploit intra- and inter-modal correlations.

III. Methodology

This section shall present general deployment principles of the multi-modal HSN, and then details the design of HDANet.

A. Multi-Modal Hybrid Sensing Network

To meet the need of crowd monitoring applications in large-scale scenarios, an HSN consisting of a large number of WiFi sniffers and a small number of cameras is constructed, as shown in Fig. 1. The HSN employs a flexible deployment approach tailored to local conditions, without adhering to any strict prerequisite, thereby expanding the practicality of the proposed method and reducing additional costs. For example, the HSN can be built by adding a small number of video cameras to an existing WiFi sniffer sensing network [29], or by deploying a large number of low-cost WiFi sniffers to an existing surveillance system covered by local videos.

The WiFi sniffers within the HSN can be dedicated multi-module WiFi sniffers, or commercial programmable WiFi APs. Equipped with a sniffing script, each sniffer tags its data with timestamps and regularly uploads the data to a server for further analysis. The deployment of sniffers can follow either manually uniform or automated schemes [45] to ensure reliable data collection and fulfill localization requirements for WiFi data preprocessing.

The camera(s) in the HSN, typically standard surveillance or security cameras, stream real-time video to the server via wired connections. While cameras can be positioned flexibly across the surveillance area, strategically placing them in areas of high crowd density can enhance multi-modal counting accuracy, as supported by findings from the ablation study in Section IV-F. Additionally, setting cameras to lower frame rates and resolutions can effectively reduce both data transmission and computational demands.

After completing the deployment of the HSN, it can be seen from Fig. 1 that the minor camera(s) cannot cover the entire large-scale surveillance area due to the limited coverage range. However, the PWS which has comprehensive and consistent sensing, can complement the spatial coverage gap occurred in the video surveillance. Furthermore, since the sensing data captured in the jointly covered area by two modalities comes from the same crowd, and the crowd behavior and distribution are usually relatively regular [29], thereby a certain spatial-temporal

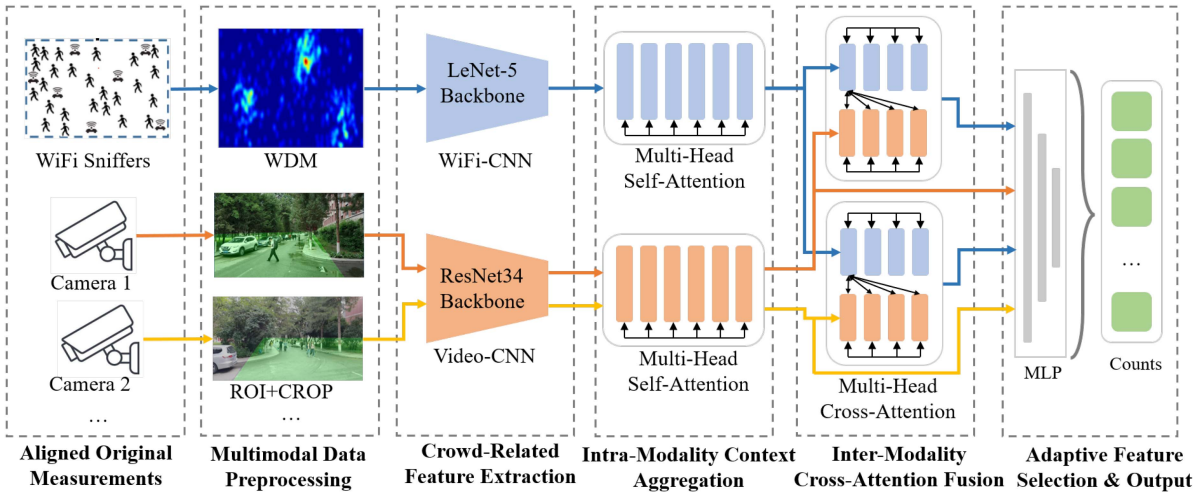

Fig. 2. Flow diagram of the proposed multi-modal crowd counting method.

correlation do exists between the global WiFi and (several) local video(s) sensing data. Consequently, utilizing cross-modality attention to learn these correlations and thus obtain cross-modality features can not only help to compensate for the deficiencies of crowd counting driven by any unimodal data, but also calibrate the WiFi sensing results in areas missed by video(s).

B. The Proposed Multi-Modal Crowd Counting Method

This section explains how to leverage the multi-modal data captured by the HSN for accurately crowd counting.

1) Overview: The proposed multi-modal crowd counting method mainly can be divided into two stages, as shown in Fig. 2.

- *Multi-modal Data Aligning and Preprocessing:* First, to overcome the randomness of PWS and meet the formatted requirement of input for the feature extraction module, we respectively employ the sliding window mechanism and WiFi-sensed density map (WDM) to filter WiFi sensing data within a certain period of time and then formalize it as robust representations of WiFi modality [46]. Second, to achieve temporal alignment and reduce computational overhead, we sample key frames, perform ROI selection, crop and resize the video modal data with the same step size, to obtain efficient image frames.
- *Multi-modal Feature Extraction, Optimization and Fusion Based on HDANet:* First, according to the abstract degree of crowd-related features contained in the preprocessed data of two modalities, CNNs with different architectures are adopted to extract the respective crowd-related features. Second, to fully exploit the intra-modal contextual correlation, the self-attention mechanism is utilized to optimize the intra-modality features. Third, the cross-modality attention mechanism and FFN are employed to match and generalize the inter-modality relationships. Finally, MLP is adopted to achieve adaptive selection of self-attention and cross-attention features, and output local and/or global crowd counts.

2) Multi-Modal Data Preprocessing: We shall describe the data preprocessing methods of two modalities in the following, respectively.

- *WiFi Modality Preprocessing:* Typically, the original WiFi sensing data is represented as separate items consisting of the MAC address of sensed mobile device, received signal strength (RSS), channel No., timestamp, etc., which includes a large amount of redundant or crowd-irrelevant information. Meanwhile, the random number of items in a time window leads to the uncertainty of inputted dimension, resulting in the difficulty of uniformly processing by the model. As such, we employ WDM to ensure compact encoding WiFi sensing data and abandon redundant information. To convert continuous physical coordinates into discrete image coordinates in pixels, a single pixel in the WDM corresponds to a $1\ \mathrm{m} \times 1\ \mathrm{m}$ square physical plane. Let n_w be the number of sensed devices, $\hat{\mathbf{L}} = \{\hat{\mathbf{l}}_1, \hat{\mathbf{l}}_2, \ldots, \hat{\mathbf{l}}_{n_w}\}$ be the set of their location estimates by an arbitrary WiFi localization algorithm [39], and $\mathbf{P} = \{p_1, p_2, \ldots, p_{n_w}\}$ be the set of pixel set transformed from $\hat{\mathbf{L}}$, thereby the intermediate image I can be generated as follows

$$I(p) = \sum_{i=1}^{n_w} \delta(p - p_i), \tag{1}$$

where δ is the impulse function. Further, the WDM can be obtained through convolving and smoothing I with a Gaussian kernel function with the size ks, i.e.,

$$WDM(p) = I(p) * G_\sigma(p), \tag{2}$$

$$G_\sigma(p) = exp\left(-\frac{||p - p_i||^2}{2\sigma^2}\right), \tag{3}$$

where $G_\sigma(p)$ is the fixed Gaussian kernel function and σ is the standard deviation. Finally, a WDM with the size of $C_w \times H_w \times W_w$ is obtained, where $C_w = 1$, H_w, and W_w are the channel number, pixel height and width of the WDM, respectively.

- *Video Modality Preprocessing:* First, since each deployed local camera has a fixed perspective, it is easy to mark the area of interest (ROI) [47] for different cameras where pedestrians may exist, resulting in excluding redundant background information with complex textures or susceptible to environmental interference. Second, to amplify the effective crowd sensing information of the video modality and ensure that the following feature extraction module can consistently process ROI images with dimensional differences, the marked frame is further cropped into a rectangle containing the ROI region and resized to a uniform size $C_v \times H_v \times W_v$, where $C_v = 1$, H_v and W_v denote the channel number, pixel height and width of the preprocessing frame from video modality, respectively.

3) HDANet: According to locations of multi-modal fusion occurring in the network, it can be divided into three modes: data-level early fusion, feature-level middle fusion and result-level late fusion [48]. Existing studies [34] show that the fusion that occurs at higher levels (i.e., middle or late fusion) generally has better predictions. In fact, WiFi and video modalities contain different forms of global and local crowd information respectively, and the expression and description perspectives are very different. Therefore, due to the heterogeneity and non-parallelism between the original data, the early fusion will cause the homogenization effect of the model, and is hard to obtain the ideal effect; while the late fusion is essentially a simple summation or weighted average of the estimates by different modalities, and is hard to extract important cross-modality information, resulting in limited improvement of the results. As such, we design the HDANet based on middle fusion, which consists of four modules: crowd-related feature extraction, intra-modality context aggregation, inter-modality cross-attention fusion, and adaptive feature selection and output.

- *Crowd-Related Feature Extraction:* This module uses CNN models with different architectures to extract modality-independent and crowd-related features at the same levels and perspectives. To avoid complicated parameter tuning and ensure the reliability, we prefer to use existing and widely validated CNN architectures as the relatively independent feature extraction backbones of two modalities. The CNN for WiFi modality (WiFi-CNN) takes the manually extracted WDM which has a higher abstract level, as the input, such that the LeNet-5 network [49] with a shallow convolutional structure is employed; while the CNN for video modality (Video-CNN) takes the preprocessed video frame which has a more primitive representation as input, and thus we employ ResNet34 [50] with a deeper convolutional structure. Therein, the convolution kernels can represent different and specific image patterns and are used to traverse 2D video frames, compute the convolution result of each small region, and finally obtain the feature plane consisting of all local features; the pooling layer is connected after each convolutional layer to optimize parameter efficiency. The WiFi-CNN includes two convolution layers consisting of 5×5 kernels with the convolution mode of step size 2, ReLU activation and maxpooling layer, such that the WiFi feature map with the size of $C'_w \times H'_w \times W'_w = 16 \times H_w/4 \times W_w/4$ is obtained; the Video-CNN contains 5 groups of 34 residual convolution layers composed of 3×3 kernels, of which the first convolution in each group adopts the convolution mode of step size 2, and ReLU activation, such that the video feature map with the size of $C'_v \times H'_v \times W'_v = 512 \times H_v/32 \times W_v/32$ is obtained.
- *Intra-Modality Context Aggregation:* This module uses the multi-head self-attention mechanism to model the channel domain and spatial domain correlation of the WiFi and video modalities, respectively, so as to optimize the crowd-related features of each. In general, the self-attention can be described as a mapping from a query (Q) and a set of key-value pair (K-V) to the output which is a weighted sum of values and the weight matrix is determined by both the query and key. According to [16], we modify the original Transformer encoder to implement our self-attention mechanism, which consists of a multi-head self-attention sublayer and a FFN sublayer. Since the multi-modal data has been preprocessed as 2D feature planes with different channels, multi-head self-attention modules with same structure but different parameters are employed for each modality. First, for each head, the feature plane $\mathbf{F} \in \mathbf{R}^{C' \times H' \times W'}$ is mapped to three types of representation in different feature subspaces through learnable 1×1 convolution kernels, which can be expressed as

$$\mathbf{Q_F} = \mathbf{W}_i^Q * \mathbf{F}, \mathbf{K_F} = \mathbf{W}_i^K * \mathbf{F}, \mathbf{V_F} = \mathbf{W}_i^V * \mathbf{F}, \quad (4)$$

where $*$ is the convolution operation, i is the subscript of the ith head, and $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, and $\mathbf{W}_i^V$ correspond to the learnable convolution kernel of the query, key, and value, respectively, with the same dimension $C'' \times H' \times W'$. Second, we divide the feature plane in each channel into S 2D patches, and then re-assemble [51] them into embedding vectors $\mathbf{Q'_F}$, $\mathbf{K'_F}$ and $\mathbf{F'_F}$ with the dimension of $S \times \hat{C}$, where $\hat{C} = \frac{HW}{S} C''_j (j \in [Q, K, V])$, such that each embedding vector contains information in both channel and spatial domains. Third, the weight matrix is obtained using scaled dot-product attention, and thus the self-attention features of a single head can be calculated by

$$\mathbf{SA}_i \left(\mathbf{Q'_F}, \mathbf{K'_F}, \mathbf{V'_F}\right) = \text{softmax}\left(\frac{\mathbf{Q'_F} \mathbf{K'^{\mathrm{T}}_F}}{\sqrt{\hat{\mathrm{C}}}}\right) \mathbf{V'_F}. \quad (5)$$

Hence, by concatenating self-attention features of all heads and leveraging the linear transformation, the multi-head self-attention features can be obtained by

$$\mathbf{MHS}(\mathbf{F}) = concat\left(\mathbf{SA}_1, \ldots, \mathbf{SA}_i, \ldots, \mathbf{SA}_h\right) \mathbf{W}^O, \quad (6)$$

where $\mathbf{W}^O$ is the transformation matrix and h is the number of heads. Finally, to further optimize the fragmented feature expression, an FFN composed of two consistent fully connected layers is employed to aggregate multi-head feature $\mathbf{MHS}$, i.e.,

$$\mathbf{F}_{SA} = ReLU\left(\mathbf{MHS} \cdot \mathbf{W}_1 + b_1\right) \cdot \mathbf{W}_2 + b_2, \quad (7)$$

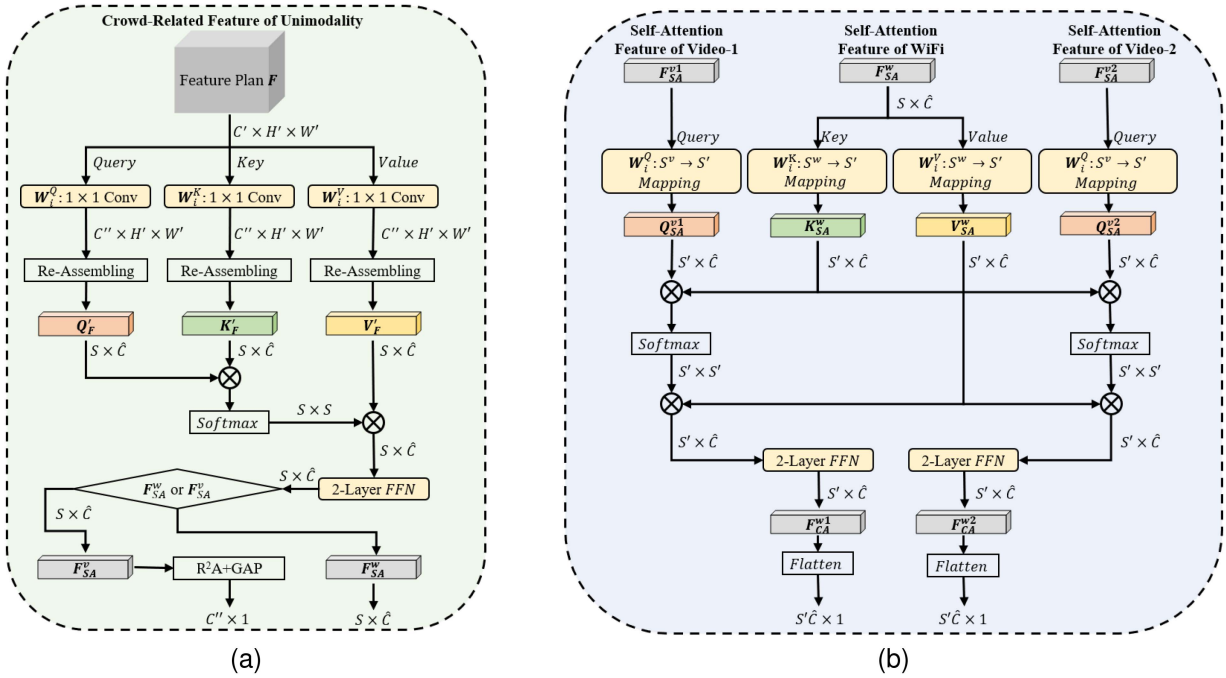

Fig. 3. Structures of multi-head attention modules. (a) Self-Attention. (b) Cross-Attention.

where $\mathbf{W}_i$ and b_i are the weight and bias of each layer, respectively, and $ReLU$ is the activation function. Meanwhile, both FNN are followed by the residual structure and layer normalization in order to fully aggregate the optimized features. Fig. 3(a) details the module structure of one head of multi-head self-attention.

- *Inter-Modality Cross-Attention Fusion:* This module uses the multi-head cross-modal attention mechanism to mine the matching relationship between modalities and propagate it to the whole feature map. Existing cross-modality attention mechanisms mainly include two modes, i.e., direct stacking features of two modalities and then utilizing self-attention to obtain cross-modality features [14], and consistently crossing between two modalities [35], both of which inevitably increase the computational overhead and are only suitable for the case that multiple modalities describe objects at different perspectives but with the same scale, e.g., the fusion of RGB image and depth image or thermal image. In contrast, we consider that the WiFi and video modalities actually describe the same crowd at different scales, thus take the global features contained in the WiFi modality as the key and value, and the local features of multiple video modalities as query vectors. Then, multiple local matching relationships are obtained by using scaled dot-product attention, and generalized and fused into the whole feature map by using the FFN. First, since the self-attention features have been re-assembled in the above module, the optimized features of two modalities can be directly used, denoted as $\mathbf{F}^w_{SA}$ and $\mathbf{F}^v_{SA}$, respectively. Second, using the similar multi-head mechanism, the linear transformation of each head to the self-attention feature of two modalities can be formulated as

$$\begin{aligned}\mathbf{Q}^v_{SA} = \mathbf{W}^Q_i * \mathbf{F}^v_{SA}, \mathbf{K}^w_{SA} = \mathbf{W}^K_i * \mathbf{F}^w_{SA}, \mathbf{V}^w_{SA} \\ = \mathbf{W}^V_i * \mathbf{F}^w_{SA}.\end{aligned} \tag{8}$$

Third, the cross-attention features of a single head and multiple head are obtained by using the similar method, which can be calculated as follows

$$\begin{aligned}&\mathbf{CA}_i\left(\mathbf{Q}^v_{SA}, \mathbf{K}^w_{SA}, \mathbf{V}^w_{SA}\right) \\ &= \text{softmax}\left(\frac{\mathbf{Q}^{\text{v}}_{\text{SA}}(\mathbf{K}^{\text{w}}_{\text{SA}})^{\text{T}}}{\sqrt{\hat{\text{C}}}}\right)\mathbf{V}^{\text{w}}_{\text{SA}}.\end{aligned} \tag{9}$$

$$\begin{aligned}&\mathbf{MHC}\left(\mathbf{F}^w_{SA}, \mathbf{F}^v_{SA}\right) \\ &= concat\left(\mathbf{CA}_1, \ldots, \mathbf{CA}_i, \ldots, \mathbf{CA}_h\right)\mathbf{W}^O.\end{aligned} \tag{10}$$

Finally, to extend the cross-modality matching relationship to the whole feature map, a similar FFN composed of two consistent fully connected layers is adopted to fuse the multi-head cross-attention feature $\mathbf{MHC}$, i.e.,

$$\mathbf{F}_{CA} = ReLU\left(\mathbf{MHC} \cdot \mathbf{W}_1 + b_1\right) \cdot \mathbf{W}_2 + b_2. \tag{11}$$

The FFN then employ residual structure and layer normalization to fully integrate cross-attention features. Fig. 3(b) details the module structure of one head of multi-head cross-attention.

- *Adaptive Feature Selection and Output:* This module adaptively selects important features from global cross-modality features and local intra-modality features by using learned weights, and thus implements the mapping to the crowd count. Since our cross-modal features are

based on the global WiFi modality, which underwent intra-modality context aggregation and cross-modality relationship matching and generalization, such that $\mathbf{F}_{CA}^{w}$ is taken as the main feature; meanwhile, considering that each video modality may provide higher contributions for its coverage when the video quality is relatively high, we adopt self-attention feature $\mathbf{F}_{SA}^{vi}$ of each video modality as the auxiliary feature. Noted that each $\mathbf{F}_{SA}^{vi}$ will be reversely re-assembled (R ^{2} A) with global average pooling (GAP) to restore the original correspondence and reduce the computation. Finally, a simple but effective 3-layers MLP is adopted to implement the adaptive selection of multi-modal features and output the count estimate, which can be synthetically formulated as

$$\hat{\mathbf{O}} = ReLU(ReLU(concat(\mathbf{F}_{CA}^{w1}, \mathbf{F}_{SA}^{v1}, \ldots, \mathbf{F}_{CA}^{wn}, \mathbf{F}_{SA}^{vn})\mathbf{W}_1 + b_1)\mathbf{W}_2 + b_2)\mathbf{W}_3 + b_3, \quad (12)$$

where $\mathbf{F}_{CA}^{wi}$ and $\mathbf{F}_{SA}^{vi}$ denote the cross-attention feature when fusing the ith local video and the self-attention feature of the ith video modality, respectively, n is the number of cameras, $\mathbf{W}_i$ and b_i are the learnable weights and biases of the ith layer of the MLP, respectively. With the adaptive selection function, this module can also effectively shield the problem of videos with low quality, namely, by reducing the weight of $\mathbf{F}_{SA}^{vi}$. Moreover, the target $\hat{\mathbf{O}}$ can be an output vector consisting of an arbitrary number of local crowds or a scalar containing only the count of global crowd, allowing the proposed method to be flexibly set based on the actual application requirements or the available crowd labels.

- *Training Method of the HDANet:* Due to the complex multi-branch structure of HDANet (e.g., the ResNet34 backbone, context aggregation module and cross-attention module) and the need for inputting multiple videos simultaneously, it is hard for conventional end-to-end training methods to optimize such a large number of learnable parameters. As such, we adopt a step-by-step pre-training approach to train each component of HDANet separately. First, after connecting the context aggregation module and output module to the WiFi-CNN and Video-CNN, they are pre-trained with WDMs and preprocessed frames as inputs, respectively. Second, the corresponding parameters of the pre-trained models are loaded and frozen into each component of the integrated HDANet, and the cross-attention and output modules are fine-tuned with pairs of synchronized WDM and preprocessed frame as inputs. Above all training processes adopt the commonly used (multi-objective) regression loss function, namely the mean square error (MSE), which can be calculated by

$$MSE\left(\hat{\mathbf{O}}_i, \mathbf{O}_i^{gt}\right) = \sum_{i=1}^{n_{tr}} \left(\hat{\mathbf{O}}_i - \mathbf{O}_i^{gt}\right)^2, \quad (13)$$

where $\hat{\mathbf{O}}_i$ and $\mathbf{O}_i^{gt}$ are the estimated and ground-truth crowd count(s) in scalar or vector, respectively, and n_{tr} is the number of training samples.

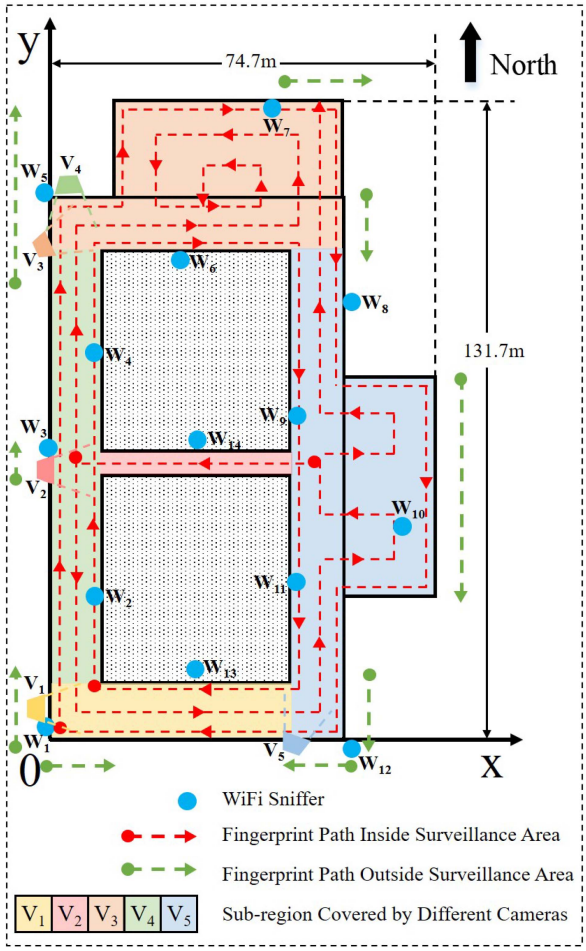

Fig. 4. The layout of the real-world dataset's testbed (Best viewed in color).

IV. Evaluation

In this section, extensive experiments are conducted on a real-world dataset to systematically and comprehensively evaluate the performance of the proposed multi-modal crowd counting method.

A. Real-World Multi-Modal Crowd Counting Dataset

A experimental HSN was deployed in a campus road network environment with a surveillance area of about 4000 m^2, and a real-world dataset was collected in the testbed, whose layout is shown in Fig. 4. In what follows, the collection and preprocessing settings of WiFi and video modalities are presented in details, respectively.

1) WiFi Sensing Data Collecting and Preprocessing: A total of 14 "Raspberry Pi 3B+" are customized as WiFi sniffers (blue circles, W_1 to W_{14}) and uniformly deployed in the surveillance area. All sniffers are strictly synchronized by embedding a high-precision time synchronization module (DS3231) and connecting a well-timed PC to calibrate the module in advance. The offline survey was conducted using 6 different smartphones, and fingerprints are generated by WiFi sensing data within a sliding time window of size $\Delta T = 60$ s and step 1 s. The path-based

fingerprint collection method [52] is employed to quickly collect fingerprints from more than 300 locations to build a location fingerprint database, and the KNN algorithm with $k = 3$ is used for localization. In the data collection stage, a total of 2280 s WiFi sensing data including a peak time after classes were obtained. As a result, 2280 WDMs with $H_w \times W_w = 140 \times 80$ are generated using a Gaussian kernel with the size of $ks = 15$ and standard deviation of $\sigma = 2$ [46].

2) Video Sensing Data Collecting and Preprocessing: A total of 5 smartphones are strictly synchronized by online calibrating their (Android) system clocks in advance, and then installed on shelves (with the height of 2.1 m) and deployed at the edge of the surveillance area (trapezoids with different colors), functioning as cameras to cover the whole area and record videos of their corresponding scenarios (see 5 sub-regions with different colors in Fig. 4). We denote the 1st to 5th scenario as S_1 to S_5 and the corresponding videos as V_1 to V_5 unless otherwise specified. To obtain the ground-truth crowd counts, each video is truncated to 2280 s length of corresponding WDMs and carefully labeled by real humans with the sampling rate of 1 frame per second, and the sum of the counts in five scenarios is taken as the total crowd count. Finally, the frame corresponding to each WDM is preprocessed by the method described in Section III-B2, and the resolution is resized to be $H_v \times W_v = 80 \times 160$ based on tradingoff between the counting accuracy and overheads (i.e., using the resolution as high as possible at an affordable time cost to improve the counting accuracy). According to the statistics of labels, the total pedestrian traffic of the dataset is 145,617, and the average pedestrian flow per second in the whole surveillance area is about 64; the relationship among average counts in each scenario is approximate to $Count(V_3) > Count(V_5) > Count(V_4) \approx Count(V_1) > Count(V_2)$; as for the video quality, V_3 suffers the problem of significant scale change, and V_4, V_5 confront serious occlusions.

B. Setup

This subsection first introduce the crowd counting baselines based on unimodality and multi-modality, respectively, and then present the training details of models and evaluation metrics.

1) Unimodal Baselines: In our proposed method, the WiFi modality data has been preprocessed into an image-like format, and a frame of the video modality is essentially a static image. Therefore, several popular CNN networks in the computer vision field, including LeNet-5 [49] (containing 2 convolutional layers), VGG11 [53] (the variant containing 8 convolutional layers), ResNet34 [50] (the variant containing 34 convolutional layers), are adopted as the feature extraction module for the unimodality. During unimodality-based counting, the transformation from classification model to regression model is achieved by slightly adjusting the output layer, as shown in Fig. 5(a); in middle fusion-based multi-modal counting, each original CNN network only retains its convolutional layers to serve as the feature extractor. Besides, the traditional global linear regression-based (denoted by G-POLY [28]) and a state-of-the-art (SOTA) deep learning-based WiFi unimodal counting methods (DNN-SCC [29]) are also implemented for comparison.

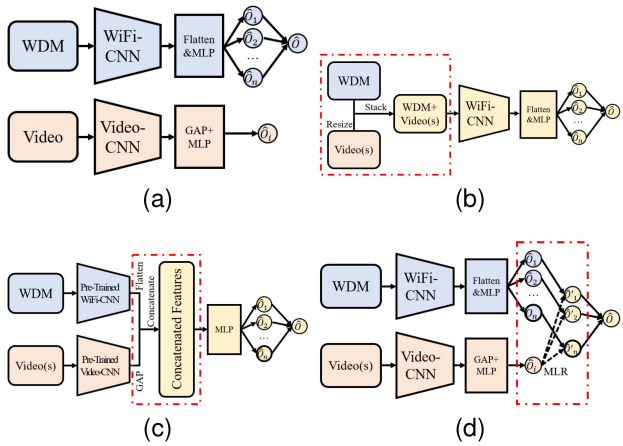

Fig. 5. Illustrations of simplified structures of unimodal and multi-modal baselines. (a) Unimodal baselines. (b) Early fusion-based multi-modal baseline. (c) Middle fusion-based multi-modal baseline. (d) Late fusion-based multi-modal baseline.

2) Multi-Modal Baselines: Due to the absence of WiFi and video-fused crowd counting methods currently, we design three different multi-modal baselines based on the commonly used fusion modes [48] in the field of MMDL. The simplified structures of each baseline are shown in Fig. 5(b)∼(d).

- *Early Fusion-Based Baseline (EF):* The fusion occurs at the input side of the model, and the feature extractor serves as the fusion network simultaneously. We adjust the frame(s) of each video into the same size as that of WDM and stacking them in the channel dimension, facilitating the processing by the subsequent model.
- *Middle Fusion-Based Baseline (MF):* The fusion occurs after feature extracting and before crowd regressing, and thus the subsequent MLP serves as the fusion network. We flatten or leverage GAP to handle the feature planes extracted from the pre-trained CNN of two modalities into one-dimensional vectors, and thus obtain the concatenated features. Since the interaction between the two modalities is not realized, the subsequent MLP can only weigh the concatenated features that are more helpful to counting.
- *Late Fusion-Based Baseline (LF):* The fusion occurs after crowd regressing, and only result-level weighting is obtained. We employ least square-based multiple linear regression (MLR) to fuse each local count of WiFi modality with the local count(s) of video modality, and the fused results are summed to obtain the total crowd count.

3) Training Details of Models: To fairly compare all methods, all models are trained with a fixed round of 500, using the MSE loss function and the Adam optimizer, where the WiFi unimodality-based baselines employ a smaller learning rate of $lr_w = 5e-4$ and the video and fusion counterparts employ a larger one $lr_v = lr_f = 1e-3$ to adapt the difference of information contained in the inputs. For HDANet and the corresponding components in baselines, the following hyperparameters are adopted: WiFi-CNN and Video-CNN employ the original setting of the corresponding CNN backbones, respectively; the multi-head self-attention modules of WiFi and video modalities

TABLE I
GLOBAL COUNTING RESULTS OF WIFI UNIMODAL AND MULTI-MODAL (WITH DIFFERENT NUMBER OF VIDEOS) METHODS

Modality	Method	MAE↓ Reduction	MSE↓	MAPE↓	ACC↑
WiFi Unimodality	G-POLY [28]	13.11	296.96	23.86%	76.14%
	DNN-SCC [29]	8.45	119.50	15.79%	84.21%
	WDM+CNN (ours)	**7.98**	**108.70**	**15.16%**	**84.84%**
WiFi+1Video Multi-modality	EF	7.15 ↓ 10.42%	89.31	14.01%	85.99%
	MF	6.82 ↓ 14.54%	75.81	**13.59%**	**86.41%**
	LF	7.15 ↓ 10.42%	88.47	13.80%	86.20%
	HDANet (ours)	**6.18 ↓ 22.61%**	**61.62**	13.73%	86.27%
WiFi+2Videos Multi-modality	EF	6.75 ↓ 15.41%	80.50	12.66%	87.34%
	MF	6.60 ↓ 17.24%	78.97	12.51%	87.49%
	LF	6.91 ↓ 13.36%	89.77	12.74%	87.26%
	HDANet (ours)	**5.89 ↓ 26.21%**	**56.06**	**12.03%**	**87.97%**

employ $h_w = 5$ heads & FFN with 512 nodes and $h_w = 16$ heads & FFN with 2,048 nodes, respectively; the multi-head cross-attention module employ $h_w = 10$ heads & FFN with 2,048 nodes; the number of nodes in each layer of the MLP is $1024/256/5$. Finally, the whole multi-modal dataset is divided into 50% training and 50% test sets, each of which contains 15 out of 30 timeslots divided from all 2280 s data [29]. Note that other ratios of dataset division were also tested, demonstrating the same or similar results, and are therefore not shown in detail.

4) Evaluation Metrics: Four metrics are adopted to comprehensively evaluate the counting performance, including two common crowd counting metrics [7], [54], namely the mean absolute error (MAE) and the MSE, and two more intuitive metrics, i.e., the mean absolute percentage error (MAPE) and the count accuracy (ACC), which are respectively defined as follows

$$MAE = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \left| \hat{\mathbf{O}}_i - \mathbf{O}_i^{gt} \right|, \tag{14}$$

$$MSE = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \left(\hat{\mathbf{O}}_i - \mathbf{O}_i^{gt} \right)^2, \tag{15}$$

$$MAPE = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \frac{\left| \hat{\mathbf{O}}_i - \mathbf{O}_i^{gt} \right|}{\mathbf{O}_i^{gt}} \times 100\%, \tag{16}$$

$$ACC = 1 - MAPE, \tag{17}$$

where n_{te} is the number of test samples, $\hat{\mathbf{O}}_i$ and $\mathbf{O}_i^{gt}$ are the estimated and ground-truth crowd counts in scenarios or the total count of the ith test sample, respectively. Generally, the MAE, MAPE, and ACC represent counting accuracy, while the MSE reflects counting stability.

C. Comparisons of Counting Performance

To validate the effectiveness of and investigate the improvement by the proposed multi-modal HDANet, we test global counting for the whole surveillance area by the SOTA and our (denoted by WDM+CNN) WiFi unimodality-based methods, and the multi-modal baselines as well as our HDANet combining with a small number of videos (i.e., 1 or 2), as shown in Table I. In addition to aforementioned metrics, the table also provides the reduced percentage of MAE achieved by multi-modal methods compared to the best WiFi unimodal baseline, i.e., our WDM+CNN. It can be clearly seen that, the counting accuracy is improved to various degrees by leveraging multi-modal methods when combining the global WiFi with local video modality, which validates the correctness of our WiFi and video-fused multi-modal counting idea. Particularly, except for the MAPE in the case of one local video combined, HDANet significantly outperforms WiFi unimodal and multi-modal baselines when combining either one local video (the MAE reduced by 22.61%, MSE reduced by 43.31%, and ACC reach 86.75%) or two local videos (the MAE reduced by 26.21%, MSE reduced by 48.43%, and ACC reach 87.97%), confirming its effectiveness in simultaneously learning intra- and inter-modality relationships of two modalities, and excellent counting accuracy.

However, it seems that the reductions of MAPE achieved by HDANet compared to other multi-modal baselines are not such obvious or even worse than the MAE, which does not comply with the common sense. In fact, the MAPE is impacted not only by the absolute error in the numerator but also by the actual count in the denominator, making it decrease faster when the actual count is smaller. As shown in Fig. 6, the test set is sorted according to the ground-truth count, and the differences between absolute errors (AEs) and between absolute percentage errors (APEs) of MF and HDANet combining with 1 video, and of LF and HDANet combining with 2 videos are plotted respectively. It can be found that HDANet significantly reduces the estimation errors when the count is large, and maintains the errors close to or better than those of baselines, where the former is more important for judging the emergency. More importantly, considering the fact that conflicts exist when the model is learning between cases of large and small counts, e.g., the model needs to have "prediction" ability due to the occlusions in large count case, but not for the small count case, it is shown that HDANet can reasonably balance such conflicts and achieve a better result.

TABLE II
BOTH THE LOCAL AND GLOBAL MAES OBTAINED BY DIFFERENT METHODS OR BACKBONES OF UNIMODALITY-BASED CROWD COUNTING

Modality	Method/Backbone	S_1	S_2	S_3	S_4	S_5	Global
WiFi Unimodality (Multi-Objective Regression)	G-POLY [28]	**2.43**	2.00	8.13	2.64	4.17	13.11
	DNN-SCC [29]	2.62	1.82	5.83	2.65	**3.28**	8.45
	LeNet-5 [49]	2.45	1.89	**5.13**	**2.48**	3.43	**7.98**
	VGG11 [53]	2.58	**1.67**	5.18	2.51	3.92	8.41
	ResNet34 [50]	2.80	1.75	6.21	2.57	4.36	9.04
Video Unimodality (Single-Objective Regression)	LeNet-5 [49]	2.37	1.19	**4.23**	2.40	4.35	8.04
	VGG11 [53]	2.14	**0.90**	6.45	2.16	3.92	9.80
	ResNet34 [50]	**2.11**	0.94	4.27	**1.81**	**3.27**	**7.17**

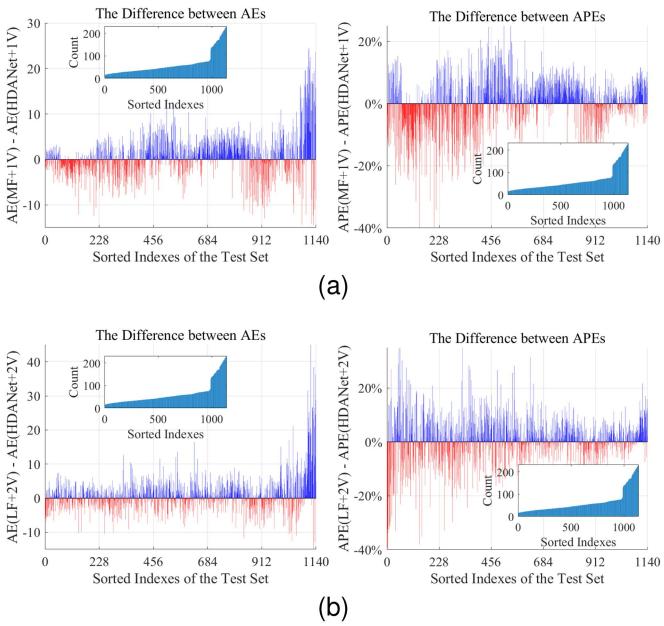

Fig. 6. The differences between AEs and between APEs of multi-modal baselines and HDANet regarding the change of actual crowd counts. (a) MF+1Video VS HDANet+1Video. (b) LF+2Videos VS HDANet+2Videos.

Since the MAE is approximately positively correlated with other metrics and can more obviously reflect performance differences, it will be employed as the main metric to investigate how several key factors affect the counting accuracy of our method in the following.

D. Effectiveness of Unimodal Branch

It is intuitive to improve the counting performance of HDANet when choosing a better handling of each modality, especially the crowd-related feature extraction module. As such, we investigate how different CNN backbones affect the counting of each modality, and the corresponding MAEs of both local (each scenario) and global (summation of all locals) counting are listed in Table II, where the WiFi unimodal method is achieved by multi-objective regression and the video unimodality, single-objective regression. It can be found that: due to the sparsity and manual feature reconstruction (i.e., constructing WDMs) of the WiFi modality, LeNet-5 which has fewer convolutional layer can achieve better counting than others, while the deeper models including VGG11 and ResNet34 are prone to the overfitting; on the contrary, the data of video modality is relatively raw and contains richer information, such that the deepest ResNet34 achieves the best result; although the smartphone-captured videos have the problem of low quality, both the local and global counting accuracy of the video modality exceeds that of the WiFi modality, which complies with the common sense. To sum up, it is reasonable for HDANet to select LeNet-5 and ResNet34 as the backbones for feature extraction of the two modalities, respectively.

E. Effectiveness of Multi-Modal Data Preprocessing

First, the effectiveness of WiFi modality preprocessing has been extensively validated in [46]. Second, in order to verify the effectiveness of our video modality preprocessing, the video unimodal baseline based on ResNet34 is utilized while keeping other settings. Four simple but effective pre-processing methods are compared, i.e., the original (ORIG), ROI-marked (ROI), our ROI-marked plus cropped (ROI+CROP), and the foreground (FG) extracted from continuous video frames [55], as illustrated in Fig. 7, and the corresponding cumulative distribution functions (CDFs) of AEs in each scenario are plotted in Fig. 8. As can be seen, ROI+CROP which has a lower computational overhead than ORIG can achieve similar (e.g., S_1 and S_3) or better (e.g., S_2, S_4 and S_5) counting accuracy by removing irrelevant background and enlarging critical areas. In addition, although the FG has the minimal lowest computation, it is more susceptible to the noise (e.g., reflections, residual shadows and slight vibrations of cameras) and occlusion, and lacks the support of image contexts, resulting in the worst results.

F. Effect of Fused Video Number

From an information theory perspective, it is expected to obtain more useful crowd-related information by fusing more videos, potentially leading to better global counting. However, in addition to adding more computational overhead, extra video data can also bring several negative effects on the fusion, including the increased noise contained in the inputted frame; the additional branches of the model leading to more parameters to be optimized simultaneously; the model faces greater difficulties

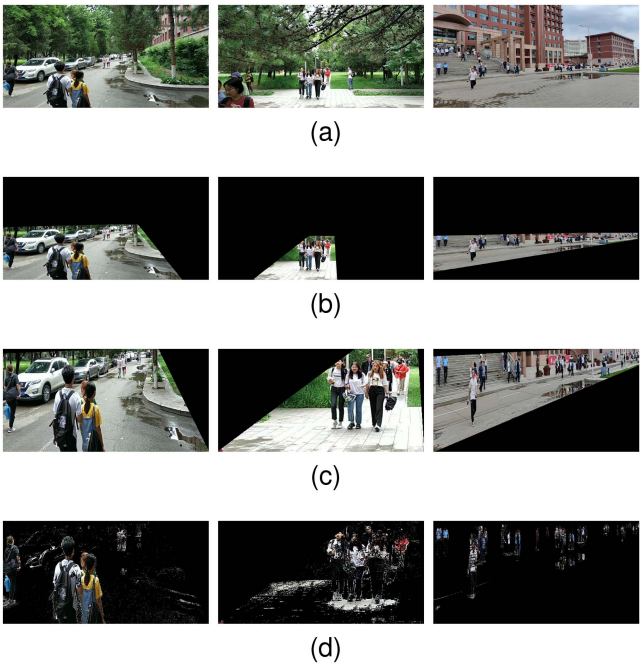

Fig. 7. Snapshots of different pre-processing methods for video modality ($V_1 \sim V_3$ from left to right). (a) ORIG. (b) ROI. (c) ROI+CROP. (d) FG.

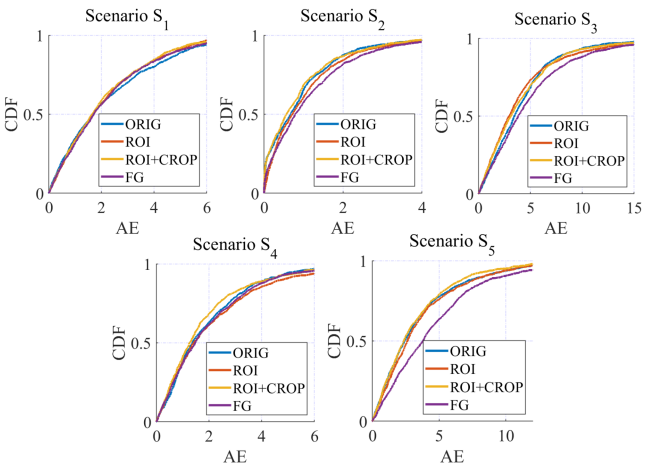

Fig. 8. CDFs of AEs obtained by video unimodal counting through different preprocessing methods in different scenarios (Best viewed in color).

in balancing the videos with different patterns of crowd distribution across different local areas, resulting in the challenge of fusing or balancing multiple local videos for better counting performance. As such, we shall investigate the effect of different numbers of fused video in the following.

To determine the selection order of different videos, a preliminary experiment is conducted to investigate the contribution or criticality of each local video. In Fig. 9, the MAEs of all multi-modal baselines combined with five different local videos are plotted. It can be found that the contribution of each local video is approximately $V_3 > V_5 > V_4 > V_1 > V_2$, positively correlating with the average pedestrian flow in each scenario,

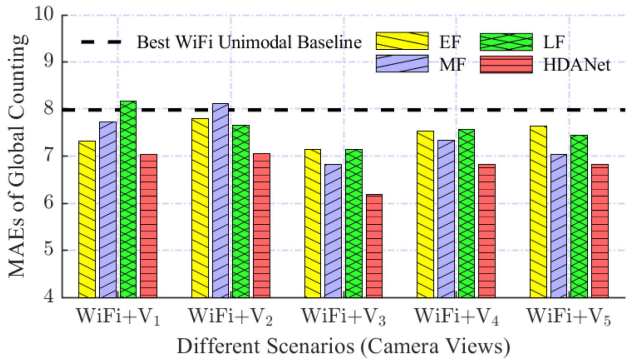

Fig. 9. Comparison of global counting MAEs between multi-modal baselines and HDANet when combining with each local video.

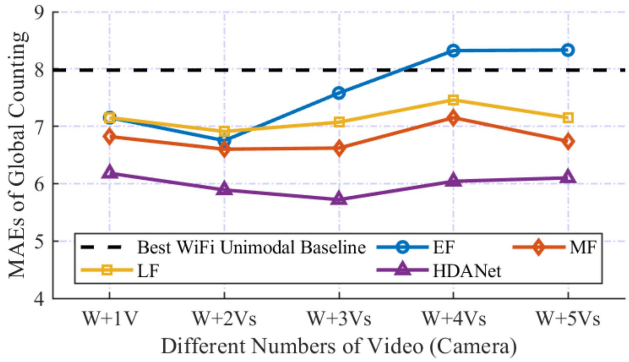

Fig. 10. MAE curves of multi-modal methods when the number of fused videos increasing (W and V are short for "WiFi" and "Video", respectively).

which implies that the deployment of cameras should prioritize the areas with higher pedestrian traffic, e.g., entrances and exits of a building, queuing or waiting areas, and locations prone to congestion.

Finally, in the case of fusing multiple videos, we adopt a greedy selection strategy based on the sorted contributions or criticalities of all videos, whose approximate optimality has been verified by exhausting all $\sum_{i=1}^{5} C_5^i$ combinations, and the MAE curves of all multi-modal methods with the number of fused videos increasing is shown in Fig. 10. As can be seen, EF employs a single feature extractor to process heterogeneous multi-modal data and suffers from the weak learning ability for handling video modalities and the inability to capture both the intra- and inter-modality relationships, resulting in sharp performance dropping when the number ≥ 3; other three methods have similar curves, but MF and LF are still limited by the insufficient capture of two relationships and also reach their performance limit when fusing two videos; HDANet overcomes above shortcomings by extracting more useful information to offset negative effects, and thus can obtain its optimal performance when fusing 3 videos and also far exceeds other baselines when fusing any number of local videos. In summary, to achieve the best performance in the practical multi-modal crowd counting, the contribution of each video should be weighed, and 2 to 3 videos with larger contributions should be selected for fusion.

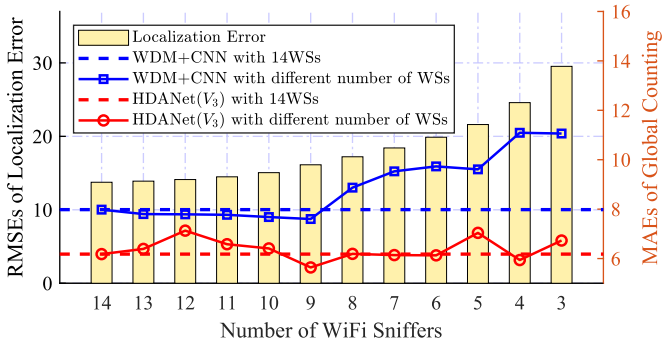

Fig. 11. The effect of the LE (RMSE) on counting error (MAE) (Best viewed in color).

TABLE III
THE GLOBAL COUNTING MAE OF HDANET WITH DIFFERENT NUMBER OF FUSED VIDEOS WHEN GRADUALLY ADDING ITS COMPONENTS

Method	W+1V	W+2Vs	W+3Vs	W+4Vs	W+5Vs
WiFi Unimodality			7.98		
M^2F	7.06	6.88	8.47	8.07	7.40
M^2F+PT	6.82	6.60	6.62	7.15	6.74
M^2F+PT+SA	6.51	6.09	5.95	6.17	6.14
HDANet (M^2F+PT +SA+CA)	**6.18**	**5.89**	**5.72**	**6.04**	**6.10**

G. Effect of Localization Errors on Counting Accuracy

Since WiFi localization plays a vital role in constructing the WDM of the WiFi modality, it is necessary to discuss how the localization error (LE) in WDM affects final crowd counting. As such, we intentionally reduce the dimensionality of the RSS fingerprints by gradually removing WiFi sniffers ($0 \sim 11$), as was done in [29], to weaken the localization ability (i.e., to increase the LEs), and plot the counting errors (MAEs) of both the optimal WiFi unimodal baseline (WDM+CNN) and multi-modal method (HDANet combining global WiFi with one local video V_3) in Fig. 11. The root mean squared error (RMSE) is utilized to measure the LE, which can be calculated by

$$RMSE = \sqrt{\frac{1}{n_{tf}} \sum_{i=1}^{n_{tf}} \left(\hat{\mathbf{l}}_i - \mathbf{l}_i^{real}\right)^2}, \tag{18}$$

where n_{tf} is the number of testing fingerprints, and $\hat{\mathbf{l}}_i$ and $\mathbf{l}_i^{real}$ are respectively the location estimates of KNN and the real location of the ith testing fingerprint's device. According to Fig. 11, we can draw the following conclusions.

First, the LEs undoubtedly lead to inaccurate crowd representations of WDMs, and thereby degrade the counting accuracy of WiFi unimodal methods. However, the counting error appears to rise when LEs are relatively large rather than uniformly scaling with the LE due to the compensations of nearby devices' localization results with small LEs [29] and the utilization of spatial correlation by neural networks [46].

Second, the slight fluctuations in the MAE of HDANet (V_3) with increasing the LE imply that the negative effect of LEs can be considerably alleviated in our multi-modal paradigm. Through investigating the difference between the WiFi unimodal and multi-modal methods, it can be intuitively explained as follows: 1) the extra and more useful pedestrian location-related information brought by the video modality is leveraged in an alternative style, i.e., the Adaptive Feature Selection Module of HDANet; 2) the errors brought by LEs in the abstract crowd features of WiFi modality are calibrated in the process of multi-modal feature interaction and matching, i.e., the Inter-Modality Cross-Attention Fusion Module of HDANet. Note that the HDANet combining global WiFi with other local videos (V_1, V_2, V_4, V_5) is also tested (but not plotted in the figure to avoid chaos), and it can be found that all the combinations are superior to that of the WiFi unimodal method but with different degrees (approximate to the ranking of each video's criticality), which conforms to above discussions. All in all, HDANet also owns the advantage of effectively combating the influence of LEs on the counting accuracy.

H. Effectiveness of Components in HDANet

In terms of model design, the superiority of HDANet mainly comes from the mode of middle fusion (M^2F), the step-by-step pre-training of each module (PT), the self-attention-based context aggregation module (SA) that captures intra-modality relationships, and the cross-attention module (CA) that captures cross-modality relationships. As such, we try to investigate their effectiveness by gradually increasing above components in the experiment of fusing different numbers of videos, as shown in Table III. We can conclude that: as the number of videos increases, the scale of the model and even the parameters to be optimized continue to grow, making the role of PT more and more obvious; SA enhances the counting accuracy in all cases with a relatively balanced amplitude by adding the context aggregation ability of each modality; CA is effective when fusing fewer local videos (≤ 3), but will lose its function when more videos are fused, which may be due to the trade-off it encounters when generalizing distinct inter-modality patterns provided by more videos. Overall, as each component is added, the counting error gradually decreases, validating the effectiveness of each component in HDANet.

To further intuitively investigate the effectiveness of SA and CA, we randomly select a multi-modal test sample, and calculate the self-attention weights of WiFi and video (V_3) modalities by $\mathrm{softmax}\left(\frac{\mathbf{Q}'_{\mathbf{F}}\mathbf{K}_{\mathbf{F}}^{\mathrm{T}}}{\sqrt{\hat{C}}}\right)$ in Section III-B3). Due to the space limit, only the part of the embedded vector of each modality is truncated for visualization, as shown in Fig. 12(a) and (b), where the weight is represented by the connection with different thicknesses. Similarly, the weights of cross-attention between WiFi modality and two video modalities (V_3 and V_5) are calculated by $\mathrm{softmax}\left(\frac{\mathbf{Q}_{\mathrm{SA}}^{\mathrm{v}}(\mathbf{K}_{\mathrm{SA}}^{\mathrm{w}})^{\mathrm{T}}}{\sqrt{\hat{C}}}\right)$ and visualized in Fig. 12(c). The observation of weight distributions reveals that: the self-attentions of two modalities different patterns of attention, i.e., the WiFi modality focuses more on its own and adjacent positions, while

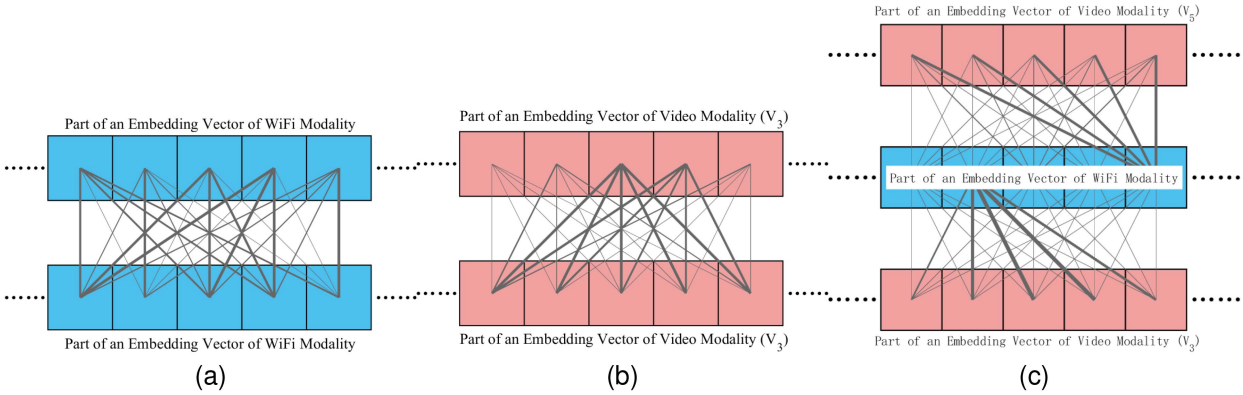

Fig. 12. Visualized self-attention (SA) and cross-attention (CA) of a randomly selected multi-modal test sample (thicknesses of connections reflect different weights). (a) Visualized SA of WiFi Modality. (b) Visualized SA of Video Modality (V_3). (c) Visualized CA between WiFi and Video Modalities (V_3 and V_5).

the video modality pays more attention to distant or important positions, which is caused by the differences in the expression of two modalities in terms of data scale and perspective; in cross-modal attention, different local videos focus on different positions of the global WiFi embedding vector, which conforms to the relationship of "local–part of global", and the relationship can be further propagated to the entire feature vector (map) through FFN. In summary, although limited by the poor interpretability of deep learning, above observations are able to confirm the validity of SA and CA to a certain extent.

I. Overhead of the Proposed Method

To validate the practicability of the proposed method, a comprehensive overhead investigation is conducted in the following. From start to finish, both the training time and inference/testing time of the preprocessing (localization and WDM construction of WiFi, ROI+CROP of the video frame), the respective pre-training or inferencing of two modalities, and the fine-tuning or inferencing of HDANet are taken into consideration. Therein, the training time refers to the time cost of pre-training or fine-tuning the corresponding models with all training samples (based on our experimental setup) on an NVIDIA A100 GPU, while the inference/testing time refers to the average time cost of the preprocessing, referencing by unimodal methods or HDANet with one test sample on one core of a common PC CPU (model Intel i5-11500H). Note that each value in the table is obtained by averaging 10 independent tests under the same condition, and the compositions of the total costs are different, i.e., the total training time equals to the sum of pre-training and fine-tuning, while the total inference/testing time equals to the sum of preprocessing of two modalities data and inferencing of HDANet.

As can be seen from Table IV, both the total training and inference/testing costs are relatively reasonable and validate the practicability of the proposed method, i.e., the training cost of HDANet is about 13+ minutes, which is very trivial in the field of deep learning, and the inference/testing cost of one multi-modal sample is less than 1 s even using the CPU, which can offer the counting frequency of 1 Hz in real applications.

TABLE IV
THE RUNTIME PERFORMANCE (UNIT *second*) OF THE PROPOSED METHOD (2 VIDEOS)

Component	Training Time	Inference/Testing Time
WiFi Preprocessing	-	4.43×10^{-2}
Video Preprocessing	-	$2.46 \times 10^{-3} \times 2$
WiFi Modality (Pre-training)	53.25	3.31×10^{-2}
Video Modality (Pre-training)	218.79×2	$4.66 \times 10^{-2} \times 2$
HDANet (Fine-tuning)	307.46	0.31
Total Time Cost	798.29	0.36

V. CONCLUSION

This paper presented an innovative multi-modal paradigm designed to enhance the accuracy and comprehensiveness of crowd counting in large-scale scenarios. In this paradigm, measurements of global WiFi and local video modalities describing the same crowd were processed with differential preprocessing, crowd-related feature extraction, context aggregation, and cross-modal matching by a unified MMDL model, namely HDANet. Extensive real-world experiments not only validated the effectiveness and superiority of our approach, but also fully investigated key challenges faced by practical applications. Moreover, the proposed method implicitly solved the problems of weak annotation and videos with low quality, where the former greatly reduces the cost of annotating video data, while the latter helps to broaden the application scenarios and reduces the requirements of computing performance. Our work provides substantial contributions to the practical application of MMDL theory and offers robust theoretical and experimental support for developing high-precision intelligent crowd monitoring systems.

In the future, we shall explore how to integrate temporal data from multiple modalities to enhance counting accuracy, incorporate cutting-edge vision techniques into HDANet's video processing branch, and develop effective synthesis methods for multi-modal data using generative deep learning models.

REFERENCES

[1] C. C. Loy, K. Chen, S. Gong, and T. Xiang, *Crowd Counting and Profiling: Methodology and Evaluation*. New York, NY, USA: Springer, 2013, pp. 347–382.

[2] V. A. Sindagi and V. M. Patel, "A survey of recent advances in CNN-based single image crowd counting and density estimation," *Pattern Recognit. Lett.*, vol. 107, pp. 3–16, 2018.

[3] X. Yu, Y. Liang, X. Lin, J. Wan, T. Wang, and H.-N. Dai, "Frequency feature pyramid network with global-local consistency loss for crowd-and-vehicle counting in congested scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9654–9664, Jul. 2022.

[4] C. I. Inc, "At least 153 dead, 133 injured in stampede during halloween festivities in Seoul," 2022. Accessed: Oct. 29, 2022. [Online]. Available: https://www.cbsnews.com/news/halloween-crowd-surge-seoul-south-korea-dozens-killed-dozens-injured/

[5] B. B. Corporation, "Indonesia: At least 125 dead in football stadium crush," 2022. Accessed: Oct. 02, 2022. [Online]. Available: https://www.bbc.com/news/world-asia-63105945

[6] J. M. Grant and P. J. Flynn, "Crowd scene understanding from video: A survey," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 2, Mar. 2017, Art. no. 19.

[7] Z. Yan et al., "Perspective-guided convolution networks for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 952–961.

[8] H. Li, E. C. L. Chan, X. Guo, J. Xiao, K. Wu, and L. M. Ni, "Wi-counter: Smartphone-based people counter using crowdsourced Wi-Fi signal data," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 4, pp. 442–452, Aug. 2015.

[9] Y. Zhao, S. Liu, F. Xue, B. Chen, and X. Chen, "DeepCount: Crowd counting with Wi-Fi using deep learning," *J. Commun. Inf. Netw.*, vol. 4, no. 3, pp. 38–52, 2019.

[10] S. Liu, Y. Zhao, and B. Chen, "WiCount: A deep learning approach for crowd counting using WiFi signals," in *Proc. IEEE Int. Symp. Parallel Distrib. Process. Appl. IEEE Int. Conf. Ubiquitous Comput. Commun.*, 2017, pp. 967–974.

[11] F.-J. Wu and G. Solmaz, "CrowdEstimator: Approximating crowd sizes with multi-modal data for Internet-of-Things services," in *Proc. 16th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, New York, NY, USA, 2018, pp. 337–349.

[12] C. Matte, M. Cunche, F. Rousseau, and M. Vanhoef, "Defeating MAC address randomization through timing attacks," in *Proc. 9th ACM Conf. Secur. Privacy Wireless Mobile Netw.*, New York, NY, USA, 2016, pp. 15–20.

[13] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[14] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10938–10947.

[15] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10494–10503.

[16] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[17] K. Chen and J.-K. Kämäräinen, "Pedestrian density analysis in public scenes with spatiotemporal tensor features," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1968–1977, Jul. 2016.

[18] A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, "Multi-object tracking through simultaneous long occlusions and split-merge conditions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 666–673.

[19] Z. Ma and A. B. Chan, "Crossing the line: Crowd counting by integer programming with local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2539–2546.

[20] X. Ding, F. He, Z. Lin, Y. Wang, H. Guo, and Y. Huang, "Crowd density estimation using fusion of multi-layer features," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 4776–4787, Aug. 2021.

[21] Q. Wang and T. P. Breckon, "Crowd counting via segmentation guided attention networks and curriculum loss," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15233–15243, Sep. 2022.

[22] A. Zhu et al., "CACrowdGAN: Cascaded attentional generative adversarial network for crowd counting," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8090–8102, Jul. 2022.

[23] X. Jiang et al., "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6126–6135.

[24] L. Liu, Z. Cao, H. Lu, H. Xiong, and C. Shen, "NSSNet: Scale-aware object counting with non-scale suppression," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3103–3114, Apr. 2022.

[25] F. Dittrich, L. E. S. de Oliveira, A. S. Britto Jr., and A. L. Koerich, "People counting in crowded and outdoor scenes using a hybrid multi-camera approach," 2017, *arXiv:1704.00326*.

[26] A. Lesani and L. Miranda-Moreno, "Development and testing of a real-time WiFi-Bluetooth system for pedestrian network monitoring, classification, and data extrapolation," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1484–1496, Apr. 2019.

[27] Y. Fukuzaki, M. Mochizuki, K. Murao, and N. Nishio, "Statistical analysis of actual number of pedestrians for Wi-Fi packet-based pedestrian flow sensing," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, New York, NY, USA, 2015, pp. 1519–1526.

[28] J. Weppner, B. Bischke, and P. Lukowicz, "Monitoring crowd condition in public spaces by tracking mobile consumer devices with WiFi interface," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.: Adjunct*, New York, NY, USA, 2016, pp. 1363–1371.

[29] L. Hao, B. Huang, B. Jia, G. Xu, and G. Mao, "Toward accurate crowd counting in large surveillance areas based on passive WiFi sensing," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14086–14096, Dec. 2023.

[30] M. S. Akhtar, D. S. Chauhan, and A. Ekbal, "A deep multi-task contextual attention framework for multi-modal affect analysis," *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 3, May 2020, Art. no. 32.

[31] J. Pan, S. Wang, and L. Fang, "Representation learning through multimodal attention and time-sync comments for affective video content analysis," in *Proc. 30th ACM Int. Conf. Multimedia*, New York, NY, USA, 2022, pp. 42–50.

[32] B. Yu, Z. Zhang, D. Zhao, and Y. Wang, "Audio-visual speech enhancement with deep multi-modality fusion," in *Proc. 5th Int. Conf. Inf. Commun. Signal Process.*, 2022, pp. 143–147.

[33] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *Proc. Eur. Symp. Artif. Neural Netw.*, 2016, pp. 509–514.

[34] J. Schlosser, C. K. Chow, and Z. Kira, "Fusing LiDAR and images for pedestrian detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 2198–2205.

[35] Y. Zhang, S. Choi, and S. Hong, "Spatio-channel attention blocks for cross-modal crowd counting," in *Proc. Eur. Conf. Comput. Vis.*, L. Wang, J. Gall, T.-J. Chin, I. Sato, and R. Chellappa, Eds., Cham, Switzerland: Springer Nature, 2023, pp. 22–40.

[36] D. Oñoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis.*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer International Publishing, 2016, pp. 615–629.

[37] S. Kumagai, K. Hotta, and T. Kurita, "Mixture of counting CNNs," *Mach. Vis. Appl.*, vol. 29, no. 7, pp. 1119–1126, Oct. 2018.

[38] X. Guo, N. Ansari, F. Hu, Y. Shao, N. R. Elikplim, and L. Li, "A survey on fusion-based indoor positioning," *IEEE Commun. Surv. Tut.*, vol. 22, no. 1, pp. 566–594, First Quarter 2020.

[39] L. Hao, B. Huang, B. Jia, and G. Mao, "DHCLoc: A device-heterogeneity-tolerant and channel-adaptive passive WiFi localization method based on DNN," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 4863–4874, Apr. 2022.

[40] T. Miyaki, T. Yamasaki, and K. Aizawa, "Multi-sensor fusion tracking using visual information and Wi-Fl location estimation," in *Proc. 1st ACM/IEEE Int. Conf. Distrib. Smart Cameras*, 2007, pp. 275–282.

[41] C. Nielsen, J. Nielsen, and V. Dehghanian, "Fusion of security camera and RSS fingerprinting for indoor multi-person tracking," in *Proc. Int. Conf. Indoor Positioning Indoor Navigation*, 2016, pp. 1–7.

[42] F. Gu, J. Niu, and L. Duan, "WAIPO: A fusion-based collaborative indoor localization system on smartphones," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2267–2280, Aug. 2017.

[43] M. D. Redžić, C. Laoudias, and I. Kyriakides, "Image and WLAN bimodal integration for indoor user localization," *IEEE Trans. Mobile Comput.*, vol. 19, no. 5, pp. 1109–1122, May 2020.

[44] C. Tang, W. Sun, X. Zhang, J. Zheng, J. Sun, and C. Liu, "A sequential-multi-decision scheme for WiFi localization using vision-based refinement," *IEEE Trans. Mobile Comput.*, vol. 23, no. 3, pp. 2321–2336, Mar. 2024.
[45] Y. Tian, B. Huang, B. Jia, and L. Zhao, "Optimizing AP and beacon placement in WiFi and BLE hybrid localization," *J. Netw. Comput. Appl.*, vol. 164, 2020, Art. no. 102673.
[46] L. Hao, B. Huang, B. Jia, and G. Mao, "On the fine-grained crowd analysis via passive WiFi sensing," *IEEE Trans. Mobile Comput.*, vol. 23, no. 6, pp. 6697–6711, Jun. 2024.
[47] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[48] J. Fayyad, M. A. Jaradat, D. Gruyer, and H. Najjaran, "Deep learning sensor fusion for autonomous vehicle perception and localization: A review," *Sensors*, vol. 20, no. 15: 4220, pp. 1–35, Jul. 2020.
[49] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[51] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 15908–15919.
[52] J. Xu et al., "Embracing spatial awareness for reliable WiFi-based indoor location systems," in *Proc. IEEE 15th Int. Conf. Mobile Ad Hoc Sensor Syst.*, 2018, pp. 281–289.
[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.
[54] Y. Fang, B. Zhan, W. Cai, S. Gao, and B. Hu, "Locality-constrained spatial transformer network for video crowd counting," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 814–819.
[55] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1999, pp. 246–252.

Lifei Hao received the BS degree in applied physics from Chongqing University, Chongqing, China, in 2012, the ME degree in computer technology, and the PhD degree in computer science and technology from Inner Mongolia University, Hohhot, China, in 2019 and 2023, respectively, where he is currently a professor with the College of Computer Science. His main research interests include Internet of Things, WiFi localization, passive WiFi sensing, and multi-modal fusion.

Baoqi Huang (Senior Member, IEEE) received the BE degree in computer science from Inner Mongolia University (IMU), Hohhot, China, in 2002, the MS degree in computer science from Peking University, Beijing, China, in 2005, and the PhD degree in information engineering from Australian National University, Canberra, ACT, Australia, in 2012. He is currently a professor with the College of Computer Science, IMU. His research interests include Internet of Things and smart sensing. He was the recipients of 2011 Annual Chinese Government Award for Outstanding Chinese Students Abroad, the Second Prize of 2022 Natural Science Award of Inner Mongolia Autonomous Region, and 2023 Annual Baosteel Outstanding Teacher Award.

Bing Jia (Member, IEEE) received the PhD degree from Jilin Univesity, Changchun, China, in 2013. She is currently an associate professor with the College of Computer Science, Inner Mongolia University, Hohhot, China. Her current research interests include indoor localization, crowdsourcing, wireless sensor networks, and mobile computing.

Guoqiang Mao (Fellow, IEEE) is a leading professor, founding director of the Research Institute of Smart Transportation, and vice-director of the ISN State Key Lab, Xidian University. Before that he was with the University of Technology Sydney and the University of Sydney. He has published 300 papers in international conferences and journals that have been cited more than 14,000 times. His H-index is 54 and is in the list of Top 2% most-cited Scientists Worldwide 2022 by Stanford University, in 2022 and 2023. He serves as a vice-director of Smart Transportation Information Engineering Society, Chinese Institute of Electronics (2022-), and was a co-chair IEEE ITS Technical Committee on Communication Networks (2014-2017). He is an editor of *IEEE Transactions on Intelligent Transportation Systems* (since 2018), *IEEE Transactions on Wireless Communications* (2014-2019), *IEEE Transactions on Vehicular Technology* (2010-2020) and received Top Editor award for outstanding contributions to the *IEEE Transactions on Vehicular Technology*, in 2011, 2014, and 2015. He has served as a chair, co-chair and TPC member in a number of international conferences. His research interest includes intelligent transport systems, Internet of Things, wireless localization techniques, wireless sensor networks, and applied graph theory and its applications in telecommunications. He is a fellow of AAIA and IET.