This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Density-Based Flow Mask Integration via Deformable Convolution for Video People Flux Estimation

Chang-Lin Wan, Feng-Kai Huang, Hong-Han Shuai
National Yang Ming Chiao Tung University
wanchunglin.eed06@g2.nctu.edu.tw, s311505011.ee11@nycu.edu.tw, hhshuai@nycu.edu.tw

Abstract

Crowd counting is currently applied in many areas, such as transportation hubs and streets. However, most of the research still focuses on counting the number of people in a single image, and there is little research on solving the problem of calculating the number of non-repeated people in a video segment. Currently, multiple object tracking is mainly relied upon for video counting, but this method is not suitable for situations where the crowd density is too high. Therefore, we propose a Flow Mask Integration Deformable Convolution network (FMDC) combined with Inter-Frame Head Contrastive Learning (IFHC) to predict the situation of people entering and exiting the screen in a density-based manner. We verify that our proposed method is highly effective in densely populated situations and diverse scenes, and the experimental results show that our proposed method surpasses existing methods.

1. Introduction

Crowd counting is widely used in various applications, including traffic monitoring [10], event management [33], and security surveillance [23, 26, 3]. In the realm of counting people in single images, multiple approaches and counting techniques have been developed [29, 22, 28, 18, 13, 43, 21, 24]. For example, in loss designation, [29, 22, 28, 18] designed more general loss functions based on optimal transport algorithm to enhance optimization accuracy. Moreover, in terms of the domain gap problem, researchers have addressed this issue by several techniques, e.g., domain-general feature extraction [13], target domain feature alignment [43], minimization of the uncertainty of the target domain [24].

Although much of the research in this field has focused on counting the number of people in a single image, there has been limited exploration into calculating the number of unique individuals in a video segment within a specific area. By solving this problem, we can facilitate a variety of applications of crowd counting beyond image-based scenarios. For instance, by calculating how many people pass through a certain area during each period, we can use this as a basis for deciding whether to open a new store rather than estimating based on the immediate number of people present. Other possible application scenarios include optimizing traffic light duration, improving road planning, and reducing congestion.

To calculate the crowd flow through a specific area, one simple solution is to aggregate counts from individual images in the video by existing crowd counting models, e.g., [18, 30, 13, 41, 32]. However, it necessitates identifying the count of repeated individuals from each frame's total count. In fact, the counting of people in video segments presents unique challenges compared to single-image counting. How to determine the accurate calculation of additional people in each subsequent frame becomes a critical aspect to consider. Previous studies have used the cross-line method to estimate the count increment by specifically focusing on pedestrians crossing a designated line. However, this approach has limitations as it fails to consider pedestrians who do not pass the line, resulting in an incomplete count. Furthermore, the cross-line method requires manual line configuration for different camera views, which is not suitable for automated crowd counting.

Another approach to address this problem is to treat it as a Multiple Object Tracking (MOT) task [39, 38, 2, 27], which tracks and maintains the identities of individuals across multiple frames. By leveraging object tracking algorithms, such as Kalman filters or deep learning-based trackers, it becomes possible to estimate the number of non-repeated individuals in a video segment. However, one of the main issues is the accuracy and reliability of object tracking algorithms, especially in dense and crowded scenes. Tracking algorithms may struggle to maintain accurate trajectories due to occlusions, interactions between individuals, abrupt changes in motion, or variations in appearance. These issues can result in ID switches, leading to a wrong flow calculation. Additionally, the computational complexity of tracking algorithms can be a concern, particularly in real-time applications where multiple individuals need to be continuously tracked across frames.

As such, the latest relevant research, DRNet [15], utilizes density maps to locate head features by identifying local maximum points. These head features are then analyzed using the optimal transport algorithm to determine the inflow and outflow numbers between different frames. However, there are still some limitations to consider. Firstly, relying solely on the density map may not always accurately extract head features, leading to potential omission or repetition of certain features. Additionally, as the number of features being compared increases, it becomes increasingly challenging to identify the outflow and inflow counts using the optimal transport algorithm based on similarity comparisons. These limitations highlight areas for further improvement and exploration in the field of individual counting.

To better address the challenges, we develop a novel approach, namely Density-Based Flow Mask Integration via Deformable Convolution (**FMDC**). First, we employ colorization self-supervised learning with the expectation that the model can initially integrate the information from two consecutive frames, while also reducing the amount of labeled data required for training. Moreover, unlike methods identifying individuals by ambiguous head features, our method involves the prediction of inflow and outflow masks, which are then multiplied with the predicted density map to predict crowd flow. As such, this approach can enhance the precision of counting based on the predicted density map directly. Additionally, it effectively addresses occlusion issues that are common in other Multiple Object Tracking (MOT) methods. Furthermore, to discern different individuals across frames, we utilize deformable convolution for spatial alignment between two frames. Since there are many positive and negative pairs in the video segment, we improve the individual discrimination capability with our proposed Inter-Frame Head Contrastive Learning (IFHC). Finally, it is important to note that video crowd datasets with tracking labels are scarce. Therefore, we created a synthetic video crowd count dataset using CARLA [12], a widely used real-world simulator in the field of autonomous driving research. By leveraging CARLA, we collect a significant amount of data to validate the superiority of our approach over other methods.

Our core contributions can be summarized as follows:

- We propose the inflow and outflow mask task for model learning, which can be applied to any density-based crowd counting method. This task aims to enhance counting accuracy and robustness by considering individuals' connection across two frames.

- We introduce the Inter-Frame Head Contrastive Learning (IFHC), which aids the model in distinguishing and counting the inflow and outflow individuals. This loss function has the potential for application in other object-based methods as well.

- We conduct experiments on two diverse and congested datasets to evaluate the performance of our proposed method. The results demonstrate that our approach outperforms strong baselines by at least 66.7% in terms of Mean Absolute Error (MAE) and 57.3% in terms of Weighted Relative Absolute Error (WRAE).

2. Related Work

2.1. Image-based Crowd Counting

Image-based crowd counting is the task of estimating the number of people in a single image. In recent years, researchers have made significant advancements in improving crowd counting techniques [21, 13, 43, 24, 30, 17, 32]. For instance, in addressing the domain gap problem, [21] tackles the issue by rescaling the images to align the head scale distribution across different datasets. [13] introduces domain-specific and domain-general modules with reconstruction and orthogonal loss to extract domain-general features for density estimation. Another approach proposed by [43] involves Point-derived Crowd Segmentation, which utilizes adversarial learning to regularize crowd density estimation in the target domain. Meanwhile, [24] applies the Shannon entropy formula to minimize the prediction uncertainty in the target domain. These methods demonstrate ongoing efforts to mitigate the domain gap problem and improve the performance of density estimation in challenging scenarios.

In addition, there are many other works on different insights of single-image crowd count. For example, [30] aims to reduce the dependency on location-wise annotations, thus introducing a dynamic counter predictor and a mixture of counter heads to achieve local-agnostic counting. [32] initially generates two cropped images from the same original image and applies the same image transformation to both. The overlapping region of the cropped images is then extracted, and self-supervised learning is performed on the density maps of these two overlapping regions. [17] designs three attention concepts, i.e., Learnable Attention Region, Local Attention Regularization and Instance Attention Loss, to enable the model to have varying levels of attention region and tolerant the spatial error of annotations. It is worth noting that we only adopt a straightforward approach by using VGG-16 as our backbone model, followed by several convolution layers, and focused on extracting precise temporal information. Therefore, combining our work with other noteworthy crowd counting approaches may lead to further improvement.

2.2. Video-based Crowd Counting

Point-based video crowd counting employs tracking methodologies such as those illustrated in [35], which blends tracking, counting, and detection techniques for drone crowd datasets. Further advances include feature warping and attention mechanisms across multiple frames [5]. Nonetheless, our approach seeks to expand the application to more crowded and complex scenes beyond drone datasets. Other strategies conceptualize the issue as a Multiple Object Tracking (MOT) task. FairMot [39], for example, uses an anchor-free object detection method and Re-ID through patch features. HeadHunter-T [27] prioritizes head feature detection and tracking. Concurrent studies integrate YOLOX [14] for object detection in MOT techniques like ByteTrack [38], which classifies detection bounding boxes into high and low-confidence associations. BoTSorT [2] compensates for camera motion in Kalman filter prediction. However, these tracking techniques can face challenges, such as ID switches issues causing overcounting and requiring processing every frame, leading to possible redundancy and increased computational time.

Another kind of approach uses cross-line methods to address the video crowd counting problem. Cross-line methods form another category of video crowd counting solutions. These involve defining a line or boundary in the scene, often a virtual line or physical barrier. Multiple lines are utilized to count people entering or exiting in some approaches [4, 7, 8], but each line counts independently, leading to inefficiency. A more efficient method was proposed by [11], who factored pedestrian velocity as line thickness, accumulating cross-line slices in patches to estimate counts. Other strategies include [20] estimating instantaneous counts on the Line of Interest (LOI) using integer programming. [40] addressed temporal slicing issues by estimating crowd counts from image pairs, leveraging density and velocity maps. [42] ensured local crowd density estimation consistency by considering spatial relations. These methods require manual line settings, limiting automation and scalability. In contrast, our method bypasses predefined lines or boundaries, using video direction to estimate crowd counts. This offers more flexibility, adaptability, and better crowd dynamics understanding, enhancing counting accuracy and efficiency.

3. Method

Fig. 1 illustrates the model architecture of FMDC. Specifically, VGG16 [19] is used as the backbone for the encoder. Subsequently, we employ three decoder branches to process the extracted features from VGG16. These three decoder branches are the color estimation branch, the density map branch, and the inflow/outflow mask branch. In the following sections, we present the problem definition, followed by each module of the proposed FMDC.

3.1. Problem Definition

Given a video segment V consisting of T frames, denoted by $\{\mathbf{I_1}, \mathbf{I_2}, \mathbf{I_3}, ..., \mathbf{I_T}\}$, where the t-th frame $\mathbf{I_t} \in \mathbb{R}^{H\times W\times 3}$ contains $N(t)$ people. Each person typically appears in a subset of consecutive frames due to the high sensing frequency of cameras. Here, our objective is to accurately count the total number of people with different identities from the first frame to the T-th frame in the video. Considering that video frames exhibit high redundancy, we divide the problem into two subproblems: estimating $N(1)$ and $N_\tau^{in}(t)$. Here, $N_\tau^{in}(t)$ represents the number of people newly entering at time t during previous τ frames, *i.e.*, from $t-\tau+1$ to t. Without loss of generality, we assume that $T = K\tau + 1$. As such, the problem can be formulated as follows:

$$\sum_{k=1}^{K} N_\tau^{in}(k\cdot\tau+1) + N(1). \tag{1}$$

In other words, our model is designed to predict the density count of the first frame $N(1)$, and subsequently predict the inflow count, $N_\tau^{in}(k\cdot\tau+1)$, from a pair of images $\mathbf{I_{(k-1)\tau+1}}$ and $\mathbf{I_{k\tau+1}}$. This prediction of $N_\tau^{in}(k\cdot\tau+1)$ continues throughout the entire video.

3.2. Self-Supervised Colorization Learning

Due to the limited availability of video crowd count datasets, we adopt a colorization pretrained learning approach inspired by [6] to mitigate the data requirements. Unlike the process in [6], which focuses on encoding and decoding a single image, our approach involves counting the inflow of people from a pair of images. Specifically, we utilize a colorful previous frame and a grayscale frame as input, expecting the model to leverage the information from the colorful image to estimate the color for the next frame.

To efficiently enable the model to share color information between frames, we consider the spatial location as the primary distinguishing factor influenced by people's movement. Drawing inspiration from the successful utilization of deformable convolution in various spatial alignment tasks such as video super-resolution [9], restoration [34], and inpainting [36], we incorporate deformable convolution layers into our model architecture. By applying deformable convolution, we aim to align individuals' positions between frames, facilitating the transfer of colorful frame features to the corresponding regions in the subsequent frame. This approach effectively promotes the efficient sharing of color information across frames. Let $\mathbf{F_t} \in \mathbb{R}^{H/4\times W/4\times C}$ denote the feature map extracted by VGG16 at the t-th frame. The whole colorization process on two frames is expressed in three steps. First, we need to generate the offset map Δp

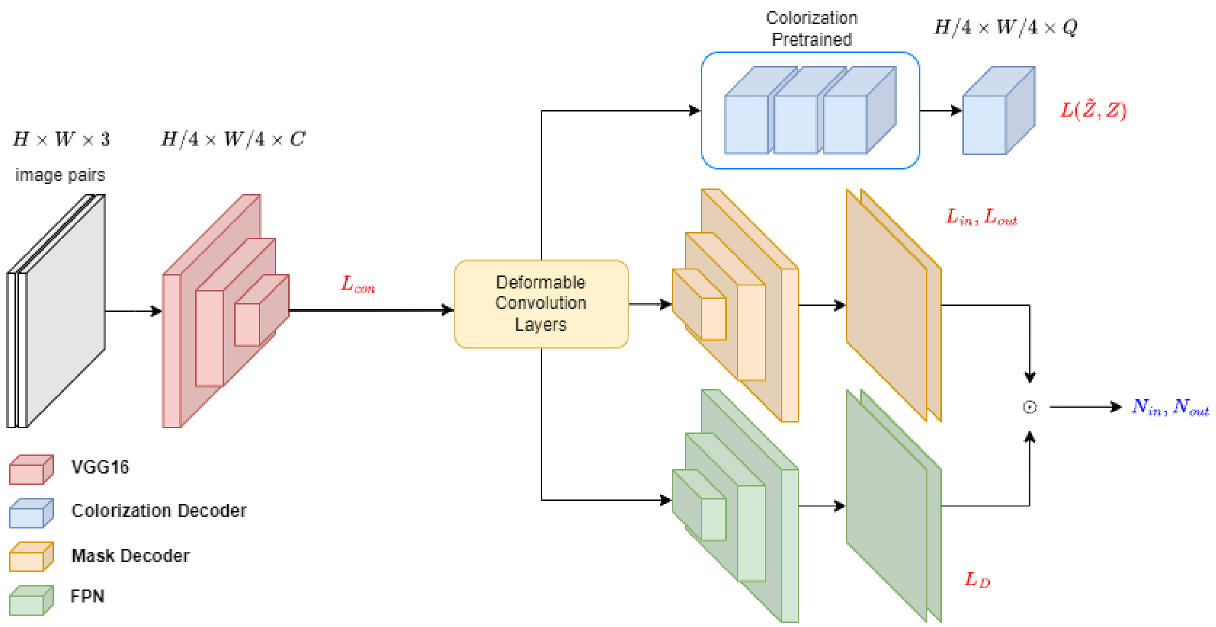

Figure 1: The training pipeline of the proposed method consists of two stages. In the first stage, we employ self-supervised colorization learning to mitigate the limited availability of training data. In the second stage, we adapt the model to predict the inflow/outflow mask and incorporate IFHC to assist the deformable convolution layers for spatial alignment. Then, the final result N_{in} and N_{out} can be obtained by multiplying the flow mask $\tilde{M}$ and the density map $\tilde{D}$.

for the deformable convolution,

$$\Delta p = G_p(concat(\mathbf{F_t}, \mathbf{F_{t+\tau}})) \tag{2}$$

where G_p and is a convolution mapping function. After having the Δp, we then apply Δp for deformable convolution layer G_d to align the spatial feature.

$$\tilde{\mathbf{F}}_\mathbf{t} = G_d(\mathbf{F_t}, \Delta p) \tag{3}$$

where $\tilde{\mathbf{F}}_\mathbf{t}$ is a spatial alignment feature computed by a deformable convolution operation. Afterwards, we concatenate the spatial alignment feature $\tilde{\mathbf{F}}_\mathbf{t}$ with $\mathbf{F_{t+\tau}}$ and then go through the colorization decoder G_c.

$$\tilde{Z} = G_c(concat(\tilde{\mathbf{F}}_\mathbf{t}, \mathbf{F_{t+\tau}})) \tag{4}$$

where $\tilde{Z} \in [0, 1]^{H/4 \times W/4 \times Q}$ represents a probability distribution of possible colors, with Q indicating the quantization level for image colors. Then, we minimize the cross entropy loss between the predicted probability $\tilde{Z}$ and ground truth color distribution Z transformed from colorful image frame $\mathbf{I_{t+\tau}}$, i.e.,

$$L(\tilde{Z}, Z) = -\sum_{p}^{HW}\sum_{q}^{Q} Z(p,q)\log \tilde{Z}(p,q). \tag{5}$$

3.3. Density Map Counting

After the colorization training, we adapt the model to perform two tasks: predicting the density map and generating the inflow/outflow mask. The inflow/outflow counts can be obtained by multiplying the density map and the inflow/outflow mask. Because of the scarcity of video crowd count data and to prevent model overfitting, we incorporate a Feature Pyramid Network (FPN) as an extension following the VGG16 backbone. Additionally, we aim to compare the performance of our inflow/outflow mask branch with DRNet, which utilizes the optimal transport for inflow/outflow count. Therefore, we adopt the same model architecture as DRNet for predicting density maps, allowing us to evaluate the effectiveness of our inflow/outflow mask branch in achieving superior performance compared to the inflow/outflow count-based approach of DRNet.

$$\tilde{D}_t, \tilde{D}_{t+\tau} = FPN(\mathbf{F_t}), FPN(\mathbf{F_{t+\tau}}). \tag{6}$$

Afterward, we minimize the MSE loss for optimizing the density map.

$$L_D = \sum_{p}^{HW}(\tilde{D}_t - D_t)^2(p) + (\tilde{D}_{t+\tau} - D_{t+\tau})^2(p). \tag{7}$$

3.4. Inflow/Outflow Mask Prediction

Similar to the colorization task, in the inflow/outflow mask task, we aim to align the individuals present in both the previous and current frames. This alignment process facilitates establishing correspondence between individuals across frames and enables accurate prediction of the inflow/outflow mask. To achieve this, we employ deformable convolution to effectively align the corresponding individuals in both frames, thereby enabling the model to discern and differentiate between different individuals across frames. The complete prediction of the inflow/outflow mask can be mathematically formulated as follows:

$$\tilde{M}_{in} = G_m(concat(\tilde{\mathbf{F}}_\mathbf{t}, \mathbf{F}_{\mathbf{t}+\tau})), \tag{8}$$

$$\tilde{M}_{out} = G_m(concat(\tilde{\mathbf{F}}_{\mathbf{t}+\tau}, \mathbf{F}_\mathbf{t})) \tag{9}$$

where $\tilde{\mathbf{F}}_\mathbf{t}$ and $\tilde{\mathbf{F}}_{\mathbf{t}+\tau}$ are computed from Eq. 2 and Eq. 3, $\tilde{M}_{in}$ and $\tilde{M}_{out} \in [0,1]^{H\times W\times 1}$ which means the probability of the in/out event. When estimating the outflow mask, we switch the input of concatenate operation, that is, $concat(\mathbf{F}_{\mathbf{t}+\tau}, \mathbf{F}_\mathbf{t})$ and the input of deformable convolution for spatial alignment is $\mathbf{F}_{\mathbf{t}+\tau}$. After predicting the density map and inflow/outflow mask, we can obtain the inflow/outflow count by multiplying the density map with the inflow/outflow mask. $\odot$ means the Hadamard product.

$$N_{in}(t+\tau) = sum(\tilde{D}_{t+\tau} \odot \tilde{M}_{in}), \tag{10}$$

$$N_{out}(t) = sum(\tilde{D}_t \odot \tilde{M}_{out}). \tag{11}$$

Afterward, the mask loss is computed by the binary cross entropy. The ground truth mask is generated by creating a square region with a value of 1 based on the center position of the person's head. The size of the region is determined by the variance of the person's head.

$$L_{in} = \sum_{p}^{HW} M_{in}(p) \cdot \tilde{M}_{in}(p) + (1 - M_{in}(p)) \cdot \tilde{M}_{in}(p), \tag{12}$$

$$L_{out} = \sum_{p}^{HW} M_{out}(p) \cdot \tilde{M}_{out}(p) + (1 - M_{out}(p)) \cdot \tilde{M}_{out}(p). \tag{13}$$

3.5. Inter-Frame Head Contrastive Learning

It is recognized that training can be unstable, and the excessive overflow of offsets for deformable convolution can significantly impact performance [34, 9]. We also find this issue when using the deformable convolution purely without any constraint. Here, we introduce IFHC for the video frame task to assist the model's spatial alignment ability. First, we extract the mutual person's head feature from extracted feature $\mathbf{F}_\mathbf{t}$, $\mathbf{F}_{\mathbf{t}+\tau}$ by the center of the bounding box. Assuming there are n mutual people, these mutual head features then form two set $\mathbf{H}^\mathbf{t} = \{f_1^t, f_2^t, ..., f_n^t\}$, $\mathbf{H}^{\mathbf{t}+\tau} = \{f_1^{t+\tau}, f_2^{t+\tau}, ..., f_n^{t+\tau}\}$ where f_n^t denotes a head feature at t-th frame. The rest of the head features form the other set $\mathbf{R} = \{f_{n+1}^t, f_{n+2}^t, ..., f_{N(t)}^t, f_{n+1}^{t+\tau}, f_{n+2}^{t+\tau}, ...f_{N(t+\tau)}^{t+\tau}\}$. Thus, the positive pairs are chosen from set $\mathbf{H}^\mathbf{t}$ and $\mathbf{H}^{\mathbf{t}+\tau}$ and the negative samples are chosen from set $\mathbf{R}$. Given the memory-intensive nature of densely crowded scenes, utilizing all negative samples from both frames may lead to excessive memory consumption. To mitigate this, we select the 50 most dissimilar negative samples relative to the desired head feature for comparison. The loss for IFHC is formulated as follows:

$$L_{con} = \sum_{i}^{n} -\log \frac{sim(\mathbf{H}^\mathbf{t}(i), \mathbf{H}^{\mathbf{t}+\tau}(i))}{\sum_{j=1}^{50} sim(\mathbf{H}^\mathbf{t}(i), \mathbf{R}_\mathbf{rank}(j))} \tag{14}$$

where the $\mathbf{R}_\mathbf{rank}$ is sorted head feature set based on similarity computed with head feature $\mathbf{H}^\mathbf{t}(i)$. The total training loss is summarized as follows:

$$L_{total} = L_D + L_{in} + L_{out} + \alpha L_{con} \tag{15}$$

where α is a hyperparameter controlling the impact of IFHC.

4. Experiment

4.1. CARLA Crowd Dataset

CARLA is an open-source simulator for autonomous driving research [12], offering a realistic virtual environment to test algorithms and systems. Given its capacity for seamlessly generating intricate street scenes, particularly those featuring pedestrians, we embark upon the task of meticulously capturing a video crowd dataset within the dynamic environment of CARLA. This comprehensive collection encompasses a diverse array of scenes, incorporating a multitude of weather conditions, varying times of day, and heterogeneous crowd distributions, among other factors.

In summary, we collect a video crowd dataset from CARLA. This dataset encompasses a total of five distinct scenes, each meticulously designed with varied people flow patterns, locations, weather conditions, and times of day. Each scene offers two different perspectives, resulting in the generation of 10 complete crowd videos within the dataset. The maximum crowd flow observed in these videos amounts to 349 individuals, while the minimum crowd flow reaches 82. The image resolution is 1920×1080. In comparison to existing synthetic crowd datasets such as GCC [31], TUB CrowdFlow [25], and CVCS [37], our proposed

Table 1: Comparisons of approaches on HT21 dataset. Underlines mean the ground truth for each scene.

Method	*val* set	Predicted counting results for each fives scene					Metrics on test set		
	CroHD01	CroHD11	CroHD12	CroHD13	CroHD14	CroHD15	MAE↓	MSE↓	WRAE↓%
	85	133	737	734	1040	321			
LOI	65.5	72.4	493.1	275.3	409.2	189.8	305.0	371.1	46.0
PHDTT	138	380	4530	5528	1531	1648	2130.4	2808.3	401.6
HeadHunter-T	145	198	636	219	458	324	253.2	351.7	32.7
FairMOT	214	144	1164	1018	632	472	256.2	300.8	44.1
ByteTrack	102	160	**761**	1467	897	460	213.2	340.2	30.8
BoT-Sort	108	174	775	2002	1000	509	315.0	574.1	46.2
DRNet	113.0	**138.4**	1017.5	623.9	659.8	**348.5**	160.7	217.3	25.1
Ours	**96.7**	138.9	664.3	**818.0**	**1005.8**	394.9	**54.2**	**61.7**	**10.7**

Table 2: Comparisons of approaches on CARLA dataset. Underlines mean the ground truth for each scene.

Method	Predicted counting results for each fives scene					Metrics on test set		
	CARLA11	CARLA12	CARLA13	CARLA14	CARLA15	MAE↓	MSE↓	WRAE↓%
	232	204	278	82	349			
ByteTrack	115	247	210	429	761	194.2	244.7	57.4
BoT-Sort	307	**235**	596	**93**	1230	260.0	417.0	59.7
DRNet	181.4	151.1	184.2	51.2	146.4	89.3	109.6	25.6
Ours	**236.0**	162.9	**241.4**	60.7	**156.6**	**59.1**	**90.0**	**20.7**

video synthetic crowd dataset stands out with its inclusion of human IDs and head bounding box annotations. This aspect makes it particularly well-suited for accurate people flux estimation and even other tasks, e.g. Crowd Tracking, Detection, Localization, Counting, etc. Additionally, our dataset exhibits remarkable diversity, encompassing multi-view perspectives, a wide range of scenes, a diverse number of counts, and varying weather and time conditions.

4.2. Experiment Setup

Evaluation metrics. We use three evaluation metrics to compare our method with baselines. The MAE and MSE are commonly used by summing all the absolute and square errors, respectively, and then taking the mean. The Weighted Relative Absolute Error (WRAE) is introduced because different video segments have different video lengths. To ensure fairness, we normalize the results by multiplying the error of each scene with the ratio of the scene length to the entire dataset. WRAE is precisely represented as follows:

$$WRAE = \sum_{i=1}^{K} \frac{\mathbf{T_i}}{\sum_{j=1}^{K} \mathbf{T_j}} \frac{|N_i - \tilde{N_i}|}{N_i} \times 100\% \qquad (16)$$

where N_i and $\tilde{N}i$ respectively represent the annotated and the estimated pedestrian number for the i-th test video. K is the total number of videos. $\mathbf{T_i}$ is the total number of frames for the i-th video.

Training details. The training details are similar to DRNet [15]. The training intervals for the frame pairs are set to 45∼80 frame intervals (about 1.8s ∼ 3.2s) and they are randomly sampled from the whole dataset. Additionally, we use the same random seed as DRNet. For data augmentation, we use the random horizontal flipping, scaling (0.8× ∼ 1.25×), and cropping (768×1024) strategies. The learning rate is initialized to $5e-5$ for VGG16 parameters, and $1e-4$ for the other parameters, and they are updated by a step decay strategy with a 0.95 rate and ADAM [16] algorithm at each epoch. The model is built upon the Pytorch framework and implemented on a TESLA V100 GPU (32G memory) with batch size 2.

4.3. Quantitative Results

Comparison method. We compare our approach with several baseline methods. The LOI method [40] utilizes the density map as a reference but leverages the velocity map to ascertain the number of individuals crossing a line. Tracking-based methods such as HeadHunter-T [27], FairMOT [39], and PHDTT [1] estimate pedestrian flow by counting individual tracks. More recent approaches like ByteTrack [38] and BoT-Sort [2] integrate the YOLOX [14] detector with their proprietary association algorithms. Among the baselines, DRNet [15] stands out as the most robust, and it is the primary method against which we benchmark our approach.

Results on two datasets. Table 1 and 2 show the predicted results on HT21 and our own CARLA datasets. In

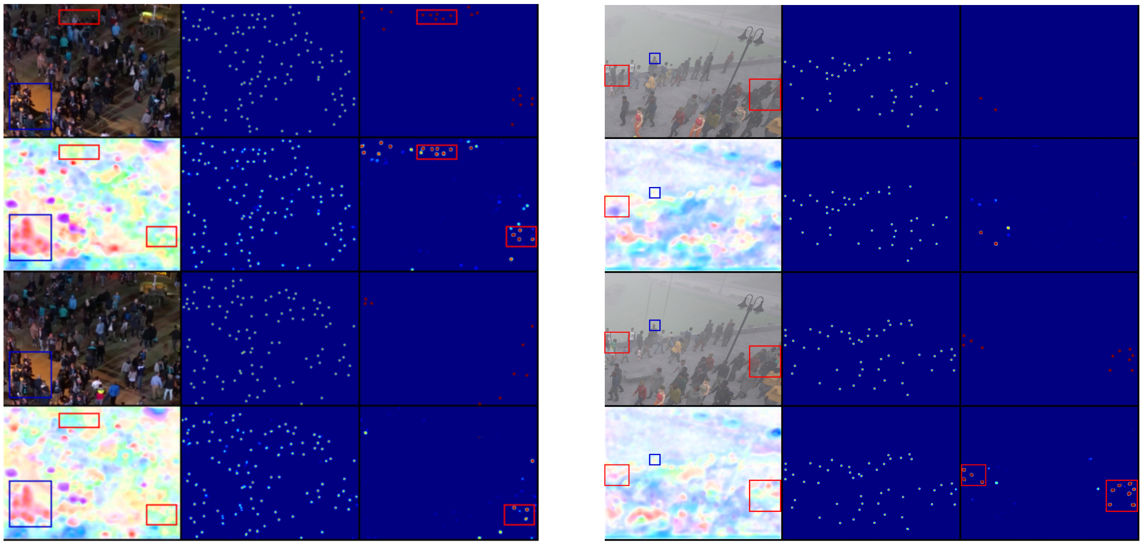

Figure 2: The counting, offset, inflow, and outflow results of our methods on the HT21 (left) and CARLA (right) dataset.

the HT21 dataset, we observe that our method performs better than other methods in complex and densely populated scenes, particularly in scenarios involving personnel entering and exiting, such as CroHD13 and CroHD14. While other methods may perform well in certain scenes, they exhibit very poor performance in others. For example, BotSort and ByteTrack achieve high MAE values of 1268 and 733, respectively, in CroHD13, indicating the occurrence of duplicate judgments due to ID switches in tracking-based approaches. On the other hand, DRNet achieves a high MAE of 280 in CroHD12. In CroHD13 and CroHD14, there are instances of underestimated predictions, which indicates the limitations of solely using head features to distinguish different individuals in dense scenes. On the contrary, our approach is comparatively more robust and generalizable than other methods. Our WRAE is only 10.7 %, which is significantly lower than that of other methods.

In the CARLA dataset, we notice that both ByteTrack and Bot-Sort also encounter issues with ID switches in CARLA15 scene. The feature comparison-based approach utilized in DRNet makes it challenging to distinguish between similar head features without considering spatial relationships. Consequently, DRNet tends to underestimate the flow of individuals. In contrast, our proposed method exhibits superior performance in densely populated scenes such as CARLA11, 13, and 15. This is attributed to the spatial alignment capability of our deformable convolution and the enhanced discrimination achieved through our designed IFHC. Furthermore, the integration of the flow mask task effectively aligns the density map, resulting in more reliable counting of inflow and outflow numbers. These factors contribute to the outstanding performance of our method across all three evaluation metrics.

Table 3: Training with different configurations on HT21.

Configurations			MAE	MSE	WRAE
Basic Conv	Deform Conv	$\mathcal{L}^{cons}$			
√			117.7	162	17.3
√		√	151	202.8	19.6
	√		169.2	245.7	21.3
	√	√	**54.2**	**61.7**	**10.7**

4.4. Ablation Study

Visualization results. Fig. 2 shows the visualization result by our model on HT21 and CARLA dataset. From top to bottom, in the first column, we have the frame $\mathbf{I_t}$, the offset map for $\tilde{\mathbf{F}}_{\mathbf{t}+\tau}$, the frame $\mathbf{I}_{\mathbf{t}+\tau}$, and the offset map for $\tilde{\mathbf{F}}_{\mathbf{t}}$. Different colors in the offset map mean the different directions of the offset. In the second column, we have D_t, $\tilde{D}_t$, $D_{t+\tau}$, $\tilde{D}_{t+\tau}$. In the third column, we have M_{out}, $\tilde{M}_{out}$, M_{in}, $\tilde{M}_{in}$. In the highlighted boxes, the red box indicates the areas where individuals enter or exit the frame, and the blue box represents areas with minimal movement of individuals. It can be observed that there are noticeable differences in the offset maps within the red box while the offset maps within the blue box appear to be more similar. This visualization result shows our deformable layer can distinguish the different people.

Training configuration. Furthermore, we verify that the combination of deformable convolution and IFHC L_{con} yields the best results. As shown in Table 3, it can be observed that the performance is worse when using only deformable convolution compared to basic convolution. However, with the addition of L_{con}, there is a significant im-

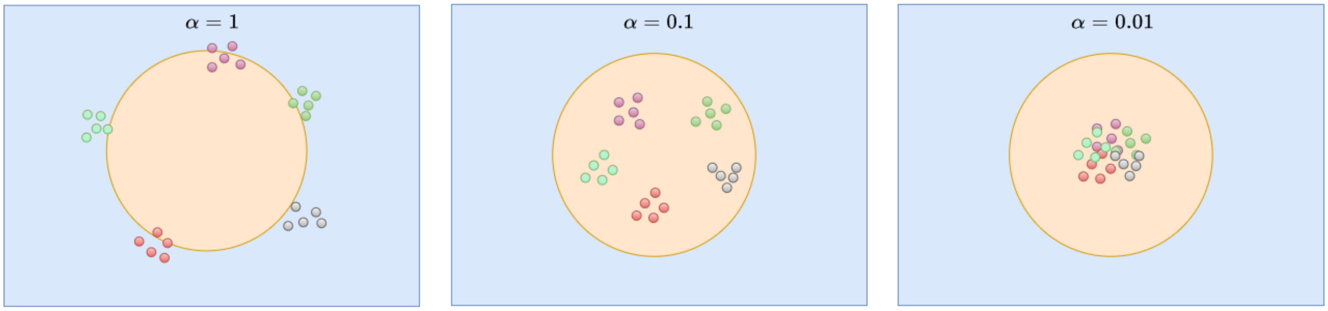

Figure 3: This is a simple illustration that demonstrates the impact of different α on the spatial distribution of head features. The orange region represents the space occupied by head features, while the blue region represents the background space. The circles of the same color represent the same individuals across different frames of the video.

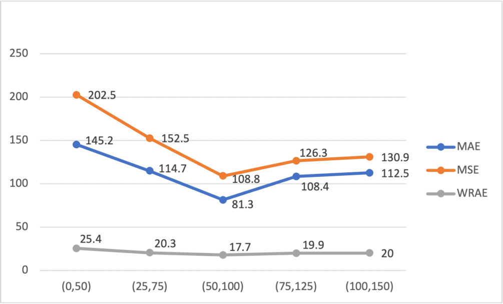

Figure 4: Optimal training intervals frame rate.

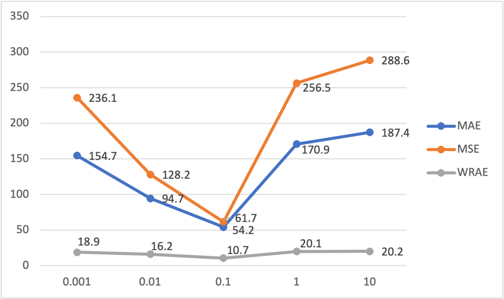

Figure 5: Hyperparameter α for IFHC.

provement in performance. Also, an interesting observation is that the basic convolution cannot incorporate L_{con} due to its lack of spatial alignment capability. Without this ability, the basic convolution is unable to effectively distinguish between different individuals across frames.

Optimal training intervals. We validate the optimal training interval on the HT21 dataset. Unlike DRNet focusing on validating the optimal testing interval, we concentrate on determining the best training interval and use it to determine the testing interval during evaluation. The reason for this approach is that in real-world applications, it is not feasible to exhaustively test the model for the best testing interval. Instead, directly applying the model with the trained interval is a more practical approach. Fig. 4 shows the different results in different frame intervals. Afterward, we fine-tuned the training interval to (40∼85) based on these results and achieved the best performance.

Hyperparameter α. Fig. 5 shows the sensitivity test on the hyperparameter α, demonstrating that the best performance is achieved when α is set to 0.1. As shown in Fig. 3, a small value of α may make the model unable to distinguish between different individuals, while a large value may cause the error of density maps in predicting all the individuals' heads. By selecting an appropriate α, all the individual head features remain within the region for head feature space and have sufficient separation from each other, enabling the model to differentiate between different individuals.

5. Conclusion

In summary, our proposed method outperforms other tracking methods and DRNet. We achieve superior results by leveraging deformable convolution for spatial alignment, allowing us to establish accurate correspondences between individuals across frames. Additionally, the incorporation of IFHC further enhances the discriminative power of our model. Furthermore, we validate our approach using the CARLA simulator, which enables us to generate realistic crowd data for video analysis. These key factors contribute to the improved performance of our method in crowd analysis tasks.

Acknowledgement

This work was supported in part by the National Science and Technology Council of Taiwan under Grants NSTC-109-2221-E-009-114-MY3, NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3.

References

[1] Xuan-thuy vo. phdtt results. [online]. .https://motchallenge.net/method/HT=7&chl=21.

[2] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Botsort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

[3] Mohammad Al-Sa'd, Serkan Kiranyaz, Iftikhar Ahmad, Christian Sundell, Matti Vakkuri, and Moncef Gabbouj. A social distance estimation and crowd monitoring system for surveillance cameras. *Sensors*, 22(2), 2022.

[4] A. Albiol, V. Naranjo, and I. Mora. Real-time high density people counter using morphological tools. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 4, pages 652–655 vol.4, 2000.

[5] Takanori Asanomi, Kazuya Nishimura, and Ryoma Bise. Multi-frame attention with feature-level warping for drone crowd tracking. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1664–1673, 2023.

[6] Haoyue Bai, Song Wen, and Shueng-Han Gary Chan. Crowd counting by self-supervised transfer colorization learning and global prior classification. *ArXiv*, abs/2105.09684, 2021.

[7] Javier Barandiaran, Berta Murguia, and Fernando Boto. Real-time people counting using multiple lines. In *2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pages 159–162, 2008.

[8] J. Bescos, J.M. Menendez, and N. Garcia. Dct based segmentation applied to a scalable zenithal people counter. In *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, volume 3, pages III–1005, 2003.

[9] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *AAAI Conference on Artificial Intelligence*, 2021.

[10] Kang Hao Cheong, Sandra Poeschmann, Joel Weijia Lai, Jin Ming Koh, U. Rajendra Acharya, Simon Ching Man Yu, and Kenneth Jian Wei Tang. Practical automated video analytics for crowd monitoring and counting. *IEEE Access*, 7:183252–183261, 2019.

[11] Yang Cong, Haifeng Gong, Song-Chun Zhu, and Yandong Tang. Flow mosaicking: Real-time pedestrian counting without scene-specific learning. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1093–1100, 2009.

[12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[13] Zhipeng Du, Jiankang Deng, and Miaojing Shi. Domain-general crowd counting in unseen scenarios. *arXiv preprint arXiv:2212.02573*, 2022.

[14] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[15] Bai Lei Gao Junyu Qi Wang Han, Tao and Ouyang Wanli. Dr.vic: Decomposition and reasoning for video individual counting. 2022.

[16] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[17] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *CVPR*, 2022.

[18] Wei Lin and Antoni B. Chan. Optimal transport minimization: Crowd localization on density maps for semi-supervised counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21663–21673, June 2023.

[19] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, 2015.

[20] Zheng Ma and Antoni B. Chan. Counting people crossing a line using integer programming and local features. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(10):1955–1969, 2016.

[21] Zhiheng Ma, Xiaopeng Hong, Xing Wei, Yunfeng Qiu, and Yihong Gong. Towards a universal model for cross-dataset crowd counting. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3185–3194, 2021.

[22] Zhiheng Ma, Xing Wei, Xiaopeng Hong, Hui Lin, Yunfeng Qiu, and Yihong Gong. Learning to count via unbalanced optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2319–2327, May 2021.

[23] Naser Hossein Motlagh, Miloud Bagaa, and Tarik Taleb. Uav-based iot platform: A crowd surveillance use case. *IEEE Communications Magazine*, 55(2):128–134, 2017.

[24] Pha Nguyen, Thanh-Dat Truong, Miaoqing Huang, Yi Liang, Ngan Le, and Khoa Luu. Self-supervised domain adaptation in crowd counting. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2786–2790, 2022.

[25] Gregory Schröder, Tobias Senst, Erik Bochinski, and Thomas Sikora. Optical flow dataset and benchmark for visual crowd analysis. In *IEEE International Conference on Advanced Video and Signals-based Surveillance*, 2018.

[26] Hongquan Song, Xuejun Liu, Xingguo Zhang, and Jiapei Hu. Real-time monitoring for crowd counting using video surveillance and gis. In *2012 2nd International Conference on Remote Sensing, Environment and Transportation Engineering*, pages 1–4, 2012.

[27] Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. Tracking pedestrian heads in dense crowd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3865–3875, June 2021.

[28] Jia Wan, Ziquan Liu, and Antoni B. Chan. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1974–1983, 2021.

[29] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. In *Advances in Neural Information Processing Systems*, 2020.

[30] Mingjie Wang, Hao Cai, Yong Dai, and Minglun Gong. Dynamic mixture of counter network for location-agnostic crowd counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 167–177, January 2023.

[31] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8198–8207, 2019.

[32] Rui Wang, Yixue Hao, Long Hu, Jincai Chen, Min Chen, and Di Wu. Self-supervised learning with data-efficient supervised fine-tuning for crowd counting. *IEEE Transactions on Multimedia*, 25:1538–1546, 2023.

[33] Tian Wang, Meina Qiao, Aichun Zhu, Guangcun Shan, and Hichem Snoussi. Abnormal event detection via the analysis of multi-frame optical flow information. *Frontiers of Computer Science*, 14, 08 2019.

[34] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[35] Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Qilong Wang, Liefeng Bo, and Siwei Lyu. Detection, tracking, and counting meets drones in crowds: A benchmark. In *CVPR*, 2021.

[36] Zhiliang Wu, Kang Zhang, Hanyu Xuan, Jian Yang, and Yan Yan. Dapc-net: Deformable alignment and pyramid context completion networks for video inpainting. *IEEE Signal Processing Letters*, 28:1145–1149, 2021.

[37] Lin Wei Zhang, Qi and Antoni B Chan. Cross-view cross-scene multi-view crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 557–567, 2021.

[38] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. 2022.

[39] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021.

[40] Zhuoyi Zhao, Hongsheng Li, Rui Zhao, and Xiaogang Wang. Crossing-line crowd counting with two-phase deep neural networks. In *European Conference on Computer Vision*, 2016.

[41] Zhiyuan Zhao and Xuelong Li. Deformable density estimation via adaptive representation. *IEEE Transactions on Image Processing*, 32:1134–1144, 2023.

[42] Huicheng Zheng, Zijian Lin, Jiepeng Cen, Zeyu Wu, and Yadan Zhao. Cross-line pedestrian counting based on spatially-consistent two-stage local crowd density estimation and accumulation. *IEEE Transactions on Circuits and Systems for Video Technology*, 29:787–799, 2019.

[43] Huilin Zhu, Jingling Yuan, Zhengwei Yang, Xian Zhong, and Zheng Wang. Fine-grained fragment diffusion for cross domain crowd counting. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 5659–5668, New York, NY, USA, 2022. Association for Computing Machinery.