DR.VIC: Decomposition and Reasoning for Video Individual Counting

Tao Han[1†],   Lei Bai[2†],   Junyu Gao[1],   Qi Wang[1*],   Wanli Ouyang[2]

[1] Northwestern Polytechnical University, Xi'an 710072, P.R. China

[2] The University of Sydney, SenseTime Computer Vision Group, Australia

hantao10200@mail.nwpu.edu.cn, {baisanshi, gjy3035, crabwq}@gmail.com,

wanli.ouyang@sydney.edu.au

Abstract

*Pedestrian counting is a fundamental tool for understanding pedestrian patterns and crowd flow analysis. Existing works (e.g., image-level pedestrian counting, cross-line crowd counting et al.) either only focus on the image-level counting or are constrained to the manual annotation of lines. In this work, we propose to conduct the pedestrian counting from a new perspective - Video Individual Counting (VIC), which counts the total number of individual pedestrians in the given video (**a person is only counted once**). Instead of relying on the Multiple Object Tracking (MOT) techniques, we propose to solve the problem by decomposing all pedestrians into the initial pedestrians who existed in the first frame and the new pedestrians with separate identities in each following frame. Then, an end-to-end Decomposition and Reasoning Network (**DRNet**) is designed to predict the initial pedestrian count with the density estimation method and reason the new pedestrian's count of each frame with the differentiable optimal transport. Extensive experiments are conducted on two datasets with congested pedestrians and diverse scenes, demonstrating the effectiveness of our method over baselines with great superiority in counting the individual pedestrians. Code: https://github.com/taohan10200/DRNet.*

1. Introduction

The world population has witnessed rapid growth, along with the accelerated urbanization. It is expected that around 70% of the world's population will live in cities [10, 51], which brings significant challenges in the city management, such as transport management, public space design, evacuation planning, and public safety. To tackle these challenges, accurately obtaining the number of pedestrians of any region in a period of time, e.g., the number of people passed

---

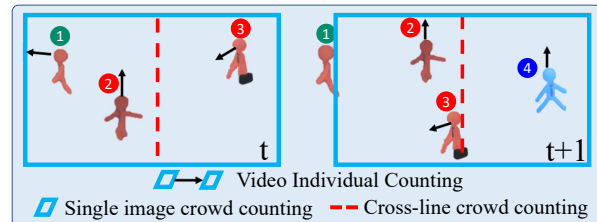† Equal contribution.
* Corresponding author.

Figure 1. Illustration of different crowd counting paradigms. Video Individual Counting count all pedestrians appearing in the video and each person is only counted once, e.g. count ①②③④ to get counting result 4. Single Image Crowd Counting targets on the image-level. Directly applying it to video individual counting will cause over-count, e.g., count pedestrians ①②③ at frame t and ②③④ at frame $t + 1$ getting result 6. Cross-line Crowd Counting targets on the video level but only focuses on the pedestrians passed the red virtual line, e.g., count ③ to get result 1.

the intersection in the past 10 minutes, is a basic problem.

Automatically estimating the pedestrian number from images/videos is a practical solution and attracts attentions of researchers from different perspectives. Specifically, Single Image Crowd Counting (SICC) [9, 18, 27, 39, 54] estimates the crowd number in the image level, which can reflect the degree of crowdedness at a certain time point. Video Crowd Counting (VCC) [15,58,59] further enhances SICC by exploring the information from the historical frames to achieve a more accurate and robust counting in the target frame. Different from SICC and VCC, the cross-line crowd counting techniques [11, 38, 63, 64] focuses on estimating the pedestrians in a period of time from videos. By manually setting a proper virtual line (e.g., the red line as illustrated in Fig. 1), cross-line crowd counting discovers the pedestrians passed this line, which could reflect the crowdness and total crowd number in the video.

Different from the settings of the works discussed above, this work targets a similar but different task –Video Individual Counting (VIC), which counts the total number of pedestrians in the given video with separate identities. As illustrated in Fig. 1, people come from all directions into the

camera view should be counted through time and each person should only be counted once. The output of VIC is the common concept regarding the total pedestrian number in a scene during a period of time (i.e., video length), which accurately measures the crowdness and popularity of an area, and thus is valuable for numerous applications, e.g., count the people attending an event, or measure the total pedestrians of an intersection.

While sharing similarities, Video Individual Counting is challenging and cannot be directly solved by existing image-level pedestrian counting and cross-line counting methods. Image level pedestrian counting methods will inevitably count the same person multiple times in adjacent frames and hugely over-estimates the pedestrian number if directly summing the number of crowd at each frame. Cross-line counting methods only count the pedestrians passing the line and thus will miss the crowd staying in the video or disjointing with the line. Besides, it requires manual annotation of the line for different camera settings, which is not in accordance with the aim of computer vision researchers in making crowd counting automated, and is also time consuming when considering a single city like London might have 500,000+ CCTV cameras.

The potential existing solution for VIC is the Multi-Object Tracking (MOT) methods [50,61], which is a general technique to identify and track all objects in the video and has been explored for specific area (e.g., counting for bus entrance [49]). The number and states of tracks in MOT can be employed to not only reveal the total number of crowd in the video, but also the inflows and outflows (the number of people getting into and out of the scene, respectively) of a period. However, since MOT is designed for tracking instead of pedestrian counting, the accuracy and efficiency of using MOT for this task would be inferior for two reasons. For accuracy, the object association in MOT depends on the detection results of multiple frames and extremely suffers from the ID switch, which heavily influence the counting performance by over-counting. For efficiency, most MOT operates on each frame, which is time-consuming and not necessary for the crowd counting task.

We propose a new paradigm for Video Individual Counting without relying on the MOT. Instead of associating all pedestrians in the whole video as MOT, we only associate each pair of frames to identify the inflow (i.e., new pedestrians) of each time slot. Specifically, we decompose all pedestrians into the initial pedestrians existed in the first frame and the new pedestrians with separate identities in the following frames. The rational behind this idea is the observation that the crowd normally stay in or pass through a region (e.g., camera view region). Only in few cases, people may pop in and out of the camera view, which is neglectable compared with the counting error. Take CroHD [50] dataset as an example, only 17 pedestrians leave and re-enter the

camera view region after an interval of 75 frames, while the total pedestrians number in the videos is 5230.

Based on this idea, an efficient and end-to-end video individual counting framework named Decomposition and Reasoning Network (DRNet) is proposed. DRNet first samples frames from the video with a time window. Then each pair of frames formed by the adjacent frames are separately fed into neural networks for generating two CNN feature maps, based on which two density maps reflecting the head locations at each frame can be predicted separately. In the next step, two sets of features containing the descriptors of each located head are sampled from the CNN feature maps and then used by a pedestrian inflow reasoning module implemented with differentiable Optimal Transport, from which the inflow (pedestrians joining the latter frame) and outflow (pedestrians leaving the former frame) of the input frame pair can be predicted. Finally, the total pedestrian count in the whole video can be obtained by integrating the crowd count of the first frame and all pedestrian inflows in the following sampled frame pairs.

Our core contributions are summarized as follows:

- We propose to decompose video individual pedestrians into the initial pedestrians and the new pedestrians at following frames, which avoids the complicated and error-prone video-level association in MOT and opens a new direction for pedestrian counting.
- We propose an efficient and end-to-end framework to achieve video individual counting, directly obtaining the new pedestrians of a frame by reasoning it with the preceding frame using differentiable optimal transport.
- Extensive experiments are conducted on two datasets covering congested crowds and diverse scenes. The experimental results demonstrate the effectiveness of the proposed methods over strong baselines.

2. Related Work

2.1. Image-level crowd counting

Image-level crowd counting refers to counting the number of people in a given static crowd image. In recent years, most state-of-the-art SICC methods concentrate on density map estimation, which integrates the density map as a count value. The CNN-based methods [5, 27, 33, 34, 57, 60, 62] show the powerful ability in feature extracting and representing than the models based on hand-crafted features [19, 31]. Apart from the density-map supervision, some methods [14, 36, 39, 55] directly exploit point-level labels to supervise counting models. Because density maps can only give a coarse count, the location distribution of people is still not available. Therefore, latest researches [2,4,16,20,32,48,53,56] target to localization for counting.

Image-Level counting from multiple frames. There are also some video-based crowd counting methods [15,29,

35, 58, 59], which exploit the temporal information to enhance the counting in the target frame.

Different to these works, our work focuses on individual pedestrian counting in the video level, which predicts the total number of the dynamic people over video frames. As discussed before, the targeted task of video individual counting cannot be solved by directly using image-level counting methods. However, since image-level counting methods can be used as a basic component in our design, the progress in image-level crowd counting can benefit our model for the sub-tasks of initial counting in the first frame and the head localization.

2.2. Video-level crowd counting

2.2.1 Tracking in crowd

Tracking in crowd [28,37] means to extract the temporal information of the crowd in a continuous sequence of images in a video. Considering the group motion behavior is consist of individual behaviour, Kratz and Nishino [23] propose a Bayesian framework that uses a space-time model for tracking individuals in a crowd. Bera *et al*. [7] propose a real-time algorithm, AdaPT, to calculate individual trajectory in a crowd scene. SSIC can be integrated to crowd tracking as well, Ren *et al*. [43] propose a tracking-by-counting method, which jointly model detection, counting, and tracking for capturing complementary information. Since tracking explicitly distinguish the identity of each person in the video, it can also be used for real-time people counting. Sun *et al*. [49] propose a RGB-D dataset that collected from the bus entrance door in surveillance view and utilize tracking to identify and count the entering and exiting people. Recently, Sundararaman [50] *et al*. propose a congested Heads Dataset (CroHD), a head detector with a Particle Filter, and a color histogram based re-identification module to track multiple people in crowded scene.

While our method also calculates cross-frame association, it does not rely on detection or video-level association. We only utilize cross-frame association to get the inflows of each time interval and integrate them together with the counting in the first frame to get the total individual number in the whole video. In this way, our design is more robust to detection and tracking errors.

2.2.2 Cross-line crowd counting

Cross-line crowd counting is a constrained direction of video pedestrian counting, which aims to count the number of pedestrian across a detection line or inside a specified region. Early works [3, 6, 8] utilize multiple lines to count the entering or existing people. These methods, however, need to perform counting independently for each line belonging to the counting zone, which is inefficient. Besides, it does not count the people who stay in the scene but do not cross the line. More importantly, it does not allow to sample the frame for reducing temporal redundancy with long interval. Cong *et al*. [11] regard pedestrians across the line as fluid flow and devise a algorithm to estimate the flow velocity field. The final count is obtained by integrating the pixels in Line of Interest (LOI) at each frame. To tackle some drawbacks in blob-centric method [11], Ma *et al*. [38] propose an integer programming approach to estimate the instantaneous counts on the LOI, which cuts the frames in a video to a set of temporal slices and then counts the people in these slice images with SICC methods. To further eliminate the limitation of the temporal slice, Zhao *et al*. [63] propose to directly estimate the crowd counts with pair of images, which resolves the LOI crowd counting by estimating the pixel-level crowd density map and crowd velocity map. The following work [64] further improves the [63] to obtain a fine-grained estimation of local crowd densities.

In general, cross-line crowd counting is limited in real application as it highly depends on the virtual line, which is hard to set for numerous videos and capture pedestrians entering and existing with random directions (e.g., squares). In contrast, our method can handle multi-direction pedestrians, and thus is applicable to more general scenes.

3. Problem Formulation

Given a video clip $\mathcal{I} = \{\mathbf{I}_0, \mathbf{I}_1, ..., \mathbf{I}_{T-1}\}$ of length T for a scene (e.g., intersection, square, exhibition), where the t-th frame $\mathbf{I}_t$ contains $N(t)$ subjects, each subject normally appears in many consecutive frames because of the relatively high sensing frequency of cameras (e.g., 25 frames per second). Our target is to count the total number of people $N(0 : T-1)$ shown in this video with distinct identities.

Instead of relying on the MOT techniques to directly obtain the crowd count with the track number at T, we propose a new solution for video individual counting. Specifically, we formulate the video individual counting problem as two sub-problems: 1) inferring crowd count $N(0)$ at the start time point, and 2) identifying the number of new identities entering the scene (inflow) at each following frame $N_{in}(t)$, which requires associating the subjects in frame $t - 1$ and t. By solving these two sub-problems, the overall video pedestrian count can be easily obtained via:

$$N(0 : T-1) = \sum_{t=1}^{t=T-1} N_{in}(t) + N(0). \qquad (1)$$

Considering that video frames are highly redundant, the inflow of most frames would be 0. Thus, we further simplify Eq. 1 by inferring the inflow every τ frames:

$$N(0 : T-1) \approx \sum_{k=1}^{k=T/\tau} N_{in}^{\tau}(k \times \tau) + N(0), \qquad (2)$$

where $N_{in}^{\tau}(t)$ is the inflow of frame $\mathbf{I}_t$ compared with $\mathbf{I}_{t-\tau}$.
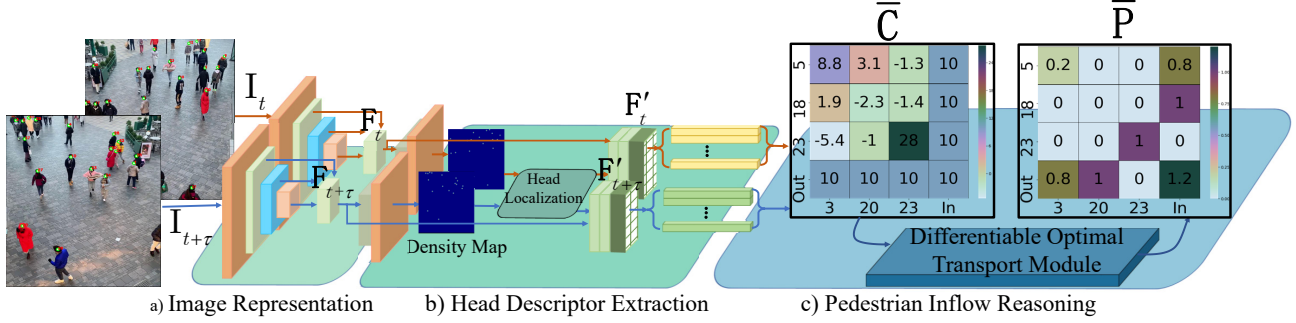
Figure 2. An overview of the DRNet architecture. DRNet is an end-to-end Video Individual Counting (VIC) framework, which takes the frames at t and $t + \tau$ as input and reasons the inflow at $t + \tau$ compared with t. Based on the image representations obtained by a shared backbone network, two density maps can be obtained to guide the extraction of head descriptors. The pedestrian inflow in pair-wise images are given by reasoning their head descriptors to $\overline{\mathbf{P}}$.

4. Method

According to the formulation defined in Sec. 3, our method should have the ability to count all pedestrians in the frame level and identify new pedestrians in a frame compared with its previous frame. We achieve this goal by designing an end-to-end network, called DRNet, to decompose video pedestrian counting as image-level pedestrian counting and cross-frame reasoning, which is composed of an image representation module, a head descriptor extraction module, and an inflow reasoning module.

Image representation. The image representation module maps two input images to feature maps in the high-level embedding space separately with a shared neural network. As elucidated in Fig. 2, we sample a pair of images $\mathbf{I}_t$ and $\mathbf{I}_{t+\tau}$ with a time interval τ from $\{\mathbf{I}_0, \mathbf{I}_1, \cdots, \mathbf{I}_{T-1}\}$. The crowd images are transformed to corresponding multi-scale feature representations $\mathbf{F}_t$ and $\mathbf{F}_{t+\tau} \in \mathbb{R}^{C \times H/4 \times W/4}$ with a backbone network, e.g., VGG-16 backbone [46] and Feature Pyramid Network (FPN) [30] in this work.

Head Descriptor Extraction. Following existing density based crowd counting works [2, 17, 20, 32, 48, 53], we use the feature maps from the image representation module to locate the head positions, which can be used to generate the descriptors (e.g., features) for each head center proposals. The details are elaborated in Sec. 4.1. The density map can also be used to generate the image level crowd count for the first frame directly in the testing phase.

Pedestrian Inflow Reasoning. Given the head descriptors of two frames from the head descriptor extraction module, the pedestrian inflow reasoning module targets on differentiation which subject is new in $\mathbf{I}_{t+\tau}$ compared with $\mathbf{I}_t$ by optimal transport. The details are elaborated in Sec. 4.2.

4.1. Head Descriptor Extraction

As shown in the middle part of Fig. 2, the head description extraction module has two branches, one is head localization branch and the other is descriptors generation
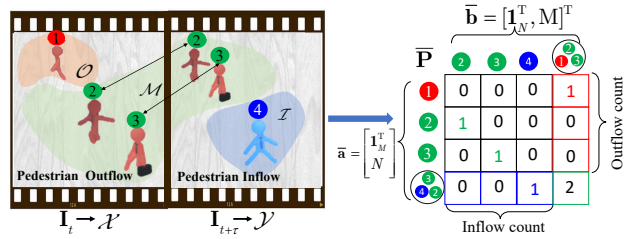

Figure 3. The sketch of the crowd transportation. With a proper solution of the transportation matrix $\overline{\mathbf{P}}$, the last row in $\overline{\mathbf{P}}$ corresponds to the **inflow** for frame $t + \tau$ that we want, *i.e.*④, the last column in $\overline{\mathbf{P}}$ describes the **outflow** regarding frame t, *i.e.* pedestrian ①, and matched pairs, *i.e.*②③.

branch. The localization branch packs several convolution and two deconvolution layers to map each image representation to a head density map, where the ground-truth head points are blurred with a Gaussian kernel $G(\sigma = 4,$ window size $= 15)$. Thus, the coordinates of the local maximums in the density map are the head center proposals.

Denote the sets of head center proposals for frame $\mathbf{I}_t$ and $\mathbf{I}_{t+\tau}$ by $\mathbf{p}_i^t := \{(h_i^t, w_i^t)\}_{i=1}^M$ and $\mathbf{p}_i^{t+\tau} := \{(h_i^{t+\tau}, w_i^{t+\tau})\}_{i=1}^N$ respectively. We first add some noises to augment these head proposals for improving the robustness,

$$\mathbf{p}_i^t \to \mathbf{p}_i^t + \mathbf{z}_1, \quad \mathbf{p}_i^{t+\tau} \to, \mathbf{p}_i^{t+\tau} + \mathbf{z}_2 \tag{3}$$

where $\mathbf{z}_2$, $\mathbf{z}_2 \sim a\mathcal{N}(0, 1)$. a controls the noise level and is empirically set as 2. Then, the augmented proposals are expanded to 8×8 regions for pooling. Finally, two sets of head descriptors are extracted with the Precise RoI Pooling [21] on the final feature map $\mathbf{F}_t'$ and $\mathbf{F}_{t+\tau}'$ (output by several convolution layers with $\mathbf{F}_t$ and $\mathbf{F}_{t+\tau}$), denoted as $\mathcal{X} := \{\mathbf{x}_1, ..., \mathbf{x}_m | \mathbf{x}_i \in \mathbb{R}^{D \times 1}\}$ for M subjects in $\mathbf{I}_t$ and $\mathcal{Y} := \{\mathbf{y}_1, ..., \mathbf{y}_n | \mathbf{y}_i \in \mathbb{R}^{D \times 1}\}$ for N subjects in $\mathbf{I}_{t+\tau}$. D is the descriptor dimension and the default setting is 256.

4.2. Pedestrian Inflow Reasoning

Target. Given the descriptors of the head center proposals obtained from the head descriptor extraction module (Sec. 4.1), the pedestrian inflow reasoning module is used for getting the inflow by associating the head center proposals. As illustrated in Fig. 3, given subjects in $\mathcal{X}$ and $\mathcal{Y}$, pedestrian inflow reasoning divides them into three categories, 1) matched instance pairs $\mathcal{M}(\subseteq \mathcal{X} \times \mathcal{Y})$ containing the people appearing in both t and $t+\tau$ frame, 2) unmatched inflow instances $\mathcal{I}(\subseteq \mathcal{Y})$ containing the people only appearing in $t + \tau$ frame, and 3) unmatched outflow instances $\mathcal{O}(\subseteq \mathcal{X})$ containing the people only appearing in t frame. Thus, during the time duration τ, the size of set $\mathcal{I}$ is the number of inflow that we want to get, *i.e.* the $N_{in}^{\tau}(t + \tau)$ in Eq. 2. The outflow set $\mathcal{O}$ is a supernumerary output.

Theoretic basis. The problem above is a typical assignment problem, for which the Hungarian algorithm [40] is a feasible solution. However, Hogarian algorithm requires a predefined distance threshold as classification basis, which would cost huge efforts to find the optimal threshold. More importantly, Hungarian algorithm is a non-differential process and only gives a hard matching (either zero or one). Hence it does not allow to optimize the image representation with the matching result. To avoid these issues, we chose the Optimal Transport (OT) theory [42] to reason this assignment problem, which is widely used to plan the transportation between two target groups under some constrains. OT also provides a differentiable association process, which makes DRNet an end-to-end trainable network.

The reasoning for inflow. Inspired by the solutions used in the graph matching and the key-point matching tasks [13, 45], the pedestrian inflow in our paper can be obtained by solving a augmented Kantorovich's OT problem,

$$\mathrm{L}_{\overline{\mathbf{C}}}(\overline{\mathbf{a}}, \overline{\mathbf{b}}) = \min_{\overline{\mathbf{P}} \in \mathbf{U}(\overline{\mathbf{a}}, \overline{\mathbf{b}})} \sum_{i \in [\![M+1]\!], j \in [\![N+1]\!]} \overline{\mathbf{C}}_{i,j} \overline{\mathbf{P}}_{i,j}, \quad (4)$$

where $\overline{\mathbf{P}}$ is a transportation matrix. As depicted in Fig. 3, its element $\overline{\mathbf{P}}_{i,j}(0 \le i{<}M, 0 \le j{<}N)$ represents the probability of the ith pedestrian in preceding frame is associated with the jth people in the current frame. Note that the augmented row $\overline{\mathbf{P}}_{(M+1,j)}(0 \le j{<}N)$ is defined to collect the pedestrian inflow (*i.e.* inflow container),

$$N_{in}^{\tau}(t + \tau) = \sum_{j=1}^{N} \overline{\mathbf{P}}_{(M+1,j)}, \quad (5)$$

Eq. 5 shows pedestrian inflow reasoning is a transportation reasoning problem, predicting the probabilities of staying in the scene and matching to the inflow/outflow. Similarly, the supernumerary outflow can be collected by the augmented column $\overline{\mathbf{P}}_{(i,N+1)}(0 \le i{<}M)$ (*i.e.* outflow container). $\overline{\mathbf{C}}$ is

the cost matrix and actually is a similarity matrix calculated with descriptor sets $\mathcal{X}$, $\mathcal{Y}$ and two augmented bins in this paper,

$$\overline{\mathbf{C}} = \left[\begin{array}{cc} \mathbf{C}_{M,N} & c_{M,1} \\ c_{1,N} & c_{1,1} \end{array} \right] \in \mathbb{R}^{(M+1)\times(N+1)}, \quad (6)$$

where $c_{M,1}, c_{1,N}$ and $c_{1,1}$ are expanded by a learnable parameter c, and act as thresholds to discriminate whether or not a person existed in both two frames. $c_{M,1}$ and $c_{1,N}$ represent the possibilities for the person matched to the outflow container and inflow container, respectively.

The (i, j)-th element $\mathbf{C}_{i,j}$ in matrix $\mathbf{C}_{M,N}$ uses the features of the i-th person at frame t and the j-th person at frame $t + \tau$ to measure their similarity as follows:

$$\mathbf{C}_{i,j} = \mathbf{x}_i^{\top} \mathbf{y}_j, \forall(i, j) \in [\![M]\!] \times [\![N]\!]. \quad (7)$$

The $\overline{\mathbf{U}}(\overline{\mathbf{a}}, \overline{\mathbf{b}})$ in Eq. 4 is a discrete measure with respect to probability vectors $\overline{\mathbf{a}}$ and $\overline{\mathbf{b}}$,

$$\overline{\mathbf{U}}(\overline{\mathbf{a}}, \overline{\mathbf{b}}) \stackrel{\text{def.}}{=} \left\{ \overline{\mathbf{P}} : \overline{\mathbf{P}} \mathbb{1}_N = \overline{\mathbf{a}} \text{ and } \overline{\mathbf{P}}^{\mathrm{T}} \mathbb{1}_M = \overline{\mathbf{b}} \right\}. \quad (8)$$

where $\mathbb{1}$ is the column vector of ones. To solve Eq. 4, we need to give a reasonable prior probability vector $\overline{\mathbf{a}}$ (for $\mathcal{X}$) and $\overline{\mathbf{b}}$ (for $\mathcal{Y}$) in advance. Actually, instances in $\mathcal{X}$ or $\mathcal{Y}$ can be regarded with the same probability to be matched as they are of equal importance in pedestrian counting. Hence their mass are all set as 1. As for inflow row and outflow column, their masses are respectively defined as N and M so as to create equal constrains. Finally, two histogram vectors $\overline{\mathbf{a}} = [\mathbb{1}_M^{\top} \ N]^{\top}$ and $\overline{\mathbf{b}} = [\mathbb{1}_N^{\top} \ M]^{\top}$ are used to solve $\overline{\mathbf{P}}$.

The overall objective of OT problem is to find matrix $\overline{\mathbf{P}}$ reasoning the pedestrians in $\mathcal{X}$ towards $\mathcal{Y}$ so that $\sum \overline{\mathbf{P}} \times \overline{\mathbf{C}}$ is the maximum. In fact, this is a linear programming problem (Eq. 4) with $N + M + 2$ constraints (Eq. 8).

Differentiable and approximate solution of OT (DOT). The final step of DRNet is to use a differentiable algorithm to solve the assignment weight matrix $\overline{\mathbf{P}}$, so that we can optimize the head descriptors with the assignment results. The standard solution of the original Kantorovich's OT problem has high time complexity and it is hard to solve. An approximated solution of the regularized Kantorovich's OT problem in [42] is given as,

$$\overline{\mathbf{P}}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j, \quad (9)$$

where $\mathbf{K}_{i,j} = e^{-\overline{\mathbf{C}}_{i,j}/\sigma}$, σ is the regularization coefficient and we set it as 1, and the vectors $\mathbf{u}$ and $\mathbf{v}$ are variables, which are solved by the Sinkhorn algorithm [12, 47]. According to the Eq. 8 and 9, we can update $\mathbf{u}$ and $\mathbf{v}$ by alternately iterating the following two equations,

$$\mathbf{u}^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}} \quad \text{and} \quad \mathbf{v}^{(\ell+1)} \stackrel{\text{def.}}{=} \frac{\mathbf{b}}{\mathbf{K}^{\mathrm{T}}\mathbf{u}^{(\ell+1)}}, \quad (10)$$

where $\mathbf{a} = \frac{\overline{\mathbf{a}}}{MN}$ and $\mathbf{b} = \frac{\overline{\mathbf{b}}}{MN}$ are, respectively, the normalized versions $\overline{\mathbf{a}}$ and $\overline{\mathbf{b}}$. $\mathbf{v}^{(0)}$ is initialized by $\mathbb{1}_N$. l is the iterations and the default setting is 100. Sarlin *et al.* [45] provide a fast-speed computation of the Sinkhorn algorithm and the time consumption is extremely low with 100 iteration, which accounts for approximately 3% training time in DRNet. Eq. 10 reveals that inferring $\overline{\mathbf{P}}$ is a completely differential process.

4.3. Loss Functions

Two loss functions are used in this framework: 1) A standard MSE loss supervises the density prediction task as widely used in the image-level crowd counting [27, 44, 62]. 2) a matching loss $\mathcal{L}_{match} = \mathcal{L}_p + \mathcal{L}_h$ can maximize the likelihood probability for positive samples and minimize the likelihood probability for hard negative samples,

$$\mathcal{L}_p = - \sum_{(i,j) \in \underset{i,j}{\arg}(\overline{\mathbf{P}}_g == 1)} \log \overline{\mathbf{P}}_{i,j}, \quad (11)$$

$$\mathcal{L}_h = - \sum_{(i,j) \in \underset{i,j}{\arg}(\overline{\mathbf{P}}_h == 1)} \log(1 - \overline{\mathbf{P}}_{i,j}), \quad (12)$$

where the $\overline{\mathbf{P}}_g \in \{0,1\}^{(m+1) \times (n+1)}$ is ground truth assignment matrix, which is generated by the the annotated association labels. $\underset{i,j}{\arg}(\overline{\mathbf{P}}_g == 1)$ returns the indexes of the elements in $\overline{\mathbf{P}}_g$ with the value as 1. Eq. 11 enforces the feature presentations of the same person to be similar in the different frames. Eq. 12 is designed to enlarge the distance between a person and his/her hard samples. Its ground truth matrix $\overline{\mathbf{P}}_h$ is adaptively generated by finding the hardest sample for each instance according to prediction $\overline{\mathbf{P}}$.

5. Experiments

5.1. Datasets

We find two benchmark datasets, i.e., CroHD [50] and SenseCrowd [25], are suitable for VIC. Both of them have annotations for head locations and associations of pedestrians. Tab. 1 describes the detailed statistics of them. These two datasets contain diverse spots, especially congested spots. For CroHD, we use four videos for training and validation, and five videos for testing as the official splitting. For SenseCrowd, all video clips are randomly divided into training (50%), validation (10%), and testing (40%). Besides, since SenseCrowd is large-scale and contains diverse scenes, we further manually label all videos with different scene labels for a more comprehensive analysis.

5.2. Evaluation Metrics

Follow existing counting tasks (e.g. crowd counting [62], vehicle counting [26]), we use Mean Absolute Error

Dataset	Videos	Frames	Head labels	Pedestrians	Time (s)
CroHD	9	11,463	2,276,838	5,230	498
SenseCrowd	634	62,938	2,344,276	43,178	12,588

Table 1. Summary of the video datasets for pedestrian counting.

(MAE) and Mean Square Error (MSE) for evaluation. Different from image-level crowd counting, we calculate them based on the whole video pedestrian count with the same person only counted once. Besides, we also use a Weighted Relative Absolute Errors (WRAE) to balance the performance on videos with different lengths and pedestrian numbers,

$$WRAE = \sum_{i=1}^{K} \frac{T_i}{\sum_{j=1}^{K} T_j} \frac{|N_i - \hat{N}_i|}{N_i} \times 100\%, \quad (13)$$

where N_i and $\hat{N}_i$ respectively represent the annotated and estimated pedestrian number for the i-th test video. K is the total number of videos. T_i is the total number of frames for the i-th video. Since our method involves association within two frames, we further use the Mean Inflow/Outflow Absolute Error (MIAE/MOAE) to reflect the association quality. (More descriptions are given in the Supplementary.)

5.3. Implementation Details

Training details: For efficient training, the time interval for each image pair is randomly sampled from range $2s \sim 8s$ to guarantee the pair contain both matched and unmatched samples. For data augmentation, we use the random horizontally flipping, scaling ($0.8\times \sim 1.25\times$), and cropping ($768 \times 1024$) strategies. The learning rate is initialized to $5e-5$ except the $1e-2$ for c, and they are updated by a step decay strategy with 0.95 rate at each epoch. Adam [22] algorithm is adopted to optimize the framework. The VGG-16 backbone is initialized with the pre-trained weights on ImageNet [24]. The model is built upon the Pytorch framework [41] and implemented on an TITAN RTX GPU (24G memory) with batchsize 4.

Testing details: In the testing phase, the time interval τ is fixed as 3s except for the time interval analysis presented in Sec. 5.5.

5.4. Overall Comparison

Comparison methods: To evaluate the effectiveness of our method, we adapt some relevant works to the individual counting task for comparison. These works are classified into two categories. 1) Tracking-based: the tracking results of three advanced MOT methods, i.e., HeadHunter-T [50], FairMOT [61], and PHDTT [52] are employed to estimate the pedestrian flow by counting their tracks. For the CroHD dataset, we first try to directly use the tracking results provided in MOTChallenge [1], but the errors (MAE/MSE) is

Methods	val set			Counting results in five testing scenes					Metrics on test set		
	CroHD01 85	MIAE↓	MOAE↓	CroHD11 133	CroHD12 737	CroHD13 734	CroHD14 1040	CroHD15 321	MAE↓	MSE↓	MRAE(%)↓
PHDTT [52]*	183	9.1	18.1	380	4530	5528	1531	1648	2130.4	2808.3	401.6
FairMOT [61]	214	6.0	6.7	**144**	1164	1018	632	472	256.2	300.8	44.1
HeadHunter-T [50]	145	**5.1**	6.2	198	636	219	458	324	253.2	351.7	32.7
LOI [63]	**65.5**	-	-	72.4	**493.1**	275.3	409.2	189.8	305.0	371.1	46.0
DRNet	113.0	6.1	**4.4**	164.6	1075.5	**752.8**	**784.5**	382.3	**141.1**	**192.3**	**27.4**

Table 2. Video individual counting performance on CroHD dataset. *: the code is not available and results come from the submission of MOTChallenge [1]. The underline fonts represent ground truth. '-' means the metrics cannot be calculated with the corresponding algorithms. Overall, DRNet obtains more accurate counts than the tracking methods [50, 52, 61] and cross-line counting method [63].

Methods	Overall					Density (for MAE)				
	MAE↓	MSE↓	MRAE(%)↓	MIAE↓	MOAE↓	D0	D1	D2	D3	D4
FairMOT [61]	35.4	62.3	48.9	4.9	4.4	13.5	22.4	67.9	84.4	145.8
HeadHunter-T [50]	30.0	50.6	38.6	4.0	4.1	11.8	25.7	56.0	92.6	131.4
LOI [63]	24.7	33.1	37.4	-	-	12.5	25.4	39.3	**39.6**	86.7
DRNet	**12.3**	**24.7**	**12.7**	**1.98**	**2.01**	**4.1**	**8.0**	**23.3**	50.0	**77.0**

Table 3. Video individual counting performance on SenseCrowd dataset. $D0 \sim D4$ respectively indicate five pedestrian density ranges: $[0, 50), [50, 100), [100, 150), [150, 200), \geq 200$. More results about the performance with different locations, day&night, indoor&outdoor are reported in Appendix.

very large. By lowering the frame rate to 0.33 FPS for Fair-MOT and 1 FPS for HeadHunter-T, trackingh-based methods produce their best results. For the SenseCrowd dataset, we use the official public code to get the tracking results (PHDTT is omitted for SenseCrowd as the code is not available). 2) Density-based: the representative cross-line crowd counting method LOI [63] is re-implemented * and assessed with our evaluation system.

Results on CroHD: To the best of our knowledge, we are the first to conduct video pedestrians counting on such a congested dataset. Tab. 2 outlines the pedestrian numbers in each scene of the testing set as well as three metrics on all videos. DRNet outperforms all tracking-based methods and cross-line method with an obvious improvement. The overall MAE and MSE are lowered to 141.1 and 192.3, respectively. The errors of some tracking methods are more than tenfold those of our MAE and MSE if we don't change its FPS in testing. The reason is that wrong associations are normal in the extremely congested scenes, which would accumulate to following associations and make tracks lose targets. In the following frames, new tracks would be added for existing pedestrians frequently. However, the reasoning error will not influence the following matching in DRNet own to the decomposition. Besides, DRNet also requires less association steps, which are only conducted in sampled frames. This also can be used to explain why DRNet surpasses other methods substantially despite the MIAE and MOAE being only slightly better than other methods. Notably, the numbers estimated by LOI [63] are all fewer than the GTs, which meets the expectation and verifies it is not a stable method to count all pedestrians in complex scenarios.

Results on SenseCrowd: Tab. 3 shows the results on SenseCrowd. DRNet makes the best predictions on the overall dataset as well as the different density subsets (D0~D4), demonstrating its effectiveness. Especially, the errors are much smaller than those in CroHD since Sense-Crowd is sparser. Our MRAE (12.7%) is relatively low, making it possible to be deployed in the future. The overall counting performance would be better if the assignment and head localization accuracies are further improved.

5.5. Ablation Study

Assignment Methods: Besides the Differential Optimal Transport (DOT) in this paper, we also consider two heuristic matching methods to achieve pedestrians association from a pair of frames: Data association in MOT [50,61] and the Hungarian algorithm [40]. In the ablation study, we first train the network with the full DRNet and then replace the DOT module with other association methods for testing. Tab. 4 shows that the association method in MOT has the largest error, whereas the Hungarian algorithm [40] can make a better matching performance with an appropriate threshold. In fact, DOT is a differential version of the Hungarian algorithm, it makes the best counting results with end-to-end learning.

Head Proposals: During training, we can use either the combination of predicted head proposals and GT points or only the GT points. Here, we analyse the contribution of the predicted head proposals. The results in the last row and the third last row in Tab. 4 show that the predicted head centers at training stage substantially improves the counting performance, decreasing MAE and MRAE by 78.0% and 26.7%, respectively.

---
*The official source code is not available.

Investigated	Settings	Counting results in five testing scenes					Metrics on test set		
		CroHD11 <u>133</u>	CroHD12 <u>737</u>	CroHD13 <u>734</u>	CroHD14 <u>1040</u>	CroHD15 <u>321</u>	MAE↓	MSE↓	MRAE(%)↓
Association methods	Tracking [50]	284	1364	1435	1917	539	526.4	604.8	87.7
	Hungarian [40]	129	421	332	331	185	313.4	395.6	45.4
Head Proposals	GT	176.9	1357.0	1118.0	1029.6	518.6	251.2	338.5	54.1
Hard Negative Pair: $\mathcal{L}_h$	W/O	151.7	1213.3	779.0	768.9	456.8	189.4	253.4	38.2
DRNet	DOT+GT+Pred+$\mathcal{L}_{match}$	164.6	1075.5	752.8	784.5	382.3	141.1	192.3	27.4

Table 4. Ablation study for DRNet. "Tracking [50]" denotes the data association method of [50]. "GT" means only using ground truth as head proposals during training. The <u>underline</u> fonts represent ground truth. All methods are with the same time interval in matching.

Hard Negative Pair Loss $\mathcal{L}_h$: Since we design $\mathcal{L}_h$ in Eq. 12 to further enlarge the feature distance of different people, thus we conduct an experiment to discuss how much gain this design brings in. Comparison between the last two rows of Tab. 4 shows that $\mathcal{L}_h$ makes a significant promotion. Take the MRAE for an example, it further drops to 27.4% by including hard negative pairs loss.

Influence of time interval: The above results of DRNet are tested with a fixed time slot τ. Here, we investigate the influence of τ to our counting performance on the CroHD dataset. Besides, we also conduct experiments on HeadHunter-T [50] at the same time intervals for comparison. As shown in Fig. 4, DRNet can achieve excellent individual counting performance with a suitable time interval (e.g., 3~4s), which also significantly decreases computation cost since less reasoning is required. However, the error rates for tracking-based HeadHunter-T steadily increase with the increase of time interval. Combined with the comparison in Tab. 2, we can conclude that DRNet can achieve much better performance when compared with the tracking-based methods in terms of both accuracy and efficiency.
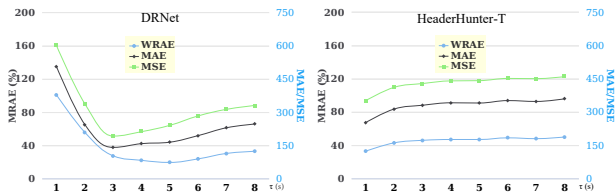

Figure 4. Errors MRAE, MAE, and MSE (Y-axis) on CroHD for different counting intervals (X-axis). DRNet achieves promising performance with a relatively large time slot (e.g., 3-4 seconds), while HeadHunter-T tends to rely on successive frames.

5.6. Qualitative Results

Fig. 5 visualizes the head proposals of the reasoned inflow outputs from DRNet on two night scenes, which is challenging for counting. Overall, DRNet makes a precise reasoning in moderately dark scenario as shown in the first row. However, there are also some failed cases in the more complicated scene as shown in Column 3 of the second row. For instance, a) the wrong assignment would decrease the pedestrian inflow number (the blue point with white box in
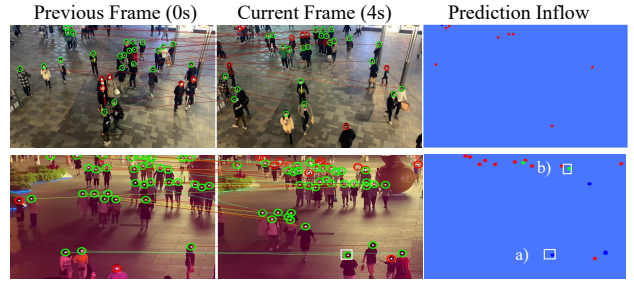

Figure 5. Visualization samples in night scenes. The green and red circles in the 1st and 2nd columns denote matched and unmatched pedestrians, respectively. The red, blue, and green points in 3rd column respectively denote correctly identified inflow, missed inflow, and over-counted new pedestrians, respectively.

last column), or b) over-claim an existing pedestrian as inflow (the green point with white box in last column).
Limitations and potential negative societal impact are disscussed in the supplementary.

6. Conclusion

We study the video individual counting task by decomposing all individuals in the video to the initial individuals at the first frame and a set of new individuals at each following frame, which is a new direction for video level crowd counting. An end-to-end learnable network named DRNet is proposed to achieve this idea by estimating the pedestrians density map and reasoning the inflows of frame-pairs with the differential optimal transport. Experiments on two datasets with congested and diverse scenes demonstrate the effectiveness of DRNet over competitive baselines. Since DRNet only reasons the association on sampled frame pairs with a large interval, the computational efficiency is also attractive. We believe this direction will make a significant promotion for crowd analysis and attract more research's interests in video individual counting and crowd analysis.

References

[1] Motchallenge. [Online]. https://motchallenge.net. 6, 7

[2] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. *arXiv preprint arXiv:2012.12482*, 2020. 2, 4

[3] Antonio Albiol, Inmaculada Mora, and Valery Naranjo. Real-time high density people counter using morphological tools. *IEEE TITS*, 2(4):204–218, 2001. 3

[4] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and Venkatesh Babu Radhakrishnan. Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE TPAMI*, 2020. 2

[5] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *CVPR*, pages 4594–4603, 2020. 2

[6] 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. IEEE, 2008. 3

[7] Aniket Bera, Nico Galoppo, Dillon Sharlet, Adam Lake, and Dinesh Manocha. Adapt: real-time adaptive pedestrian tracking for crowded scenes. In *ICRA*, pages 1801–1808. IEEE, 2014. 3

[8] Jesús Bescós, José M Menéndez, and Narciso García. Dct based segmentation applied to a scalable zenithal people counter. In *ICIP*, volume 3, pages 14–17. IEEE, 2003. 3

[9] Antoni B. Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008. 1

[10] Deevesh Chaudhary, Sunil Kumar, and Vijaypal Singh Dhaka. Video based human crowd analysis using machine learning: a survey. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–19, 2021. 1

[11] Yang Cong, Haifeng Gong, Song-Chun Zhu, and Yandong Tang. Flow mosaicking: Real-time pedestrian counting without scene-specific learning. In *CVPR*, pages 1093–1100. IEEE, 2009. 1, 3

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 26:2292–2300, 2013. 5

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, pages 224–236, 2018. 5

[14] Zihao Dong, Ruixun Zhang, Xiuli Shao, and Yumeng Li. Scale-recursive network with point supervision for crowd scene analysis. *Neurocomputing*, 384:314–324, 2020. 2

[15] Yanyan Fang, Biyun Zhan, Wandi Cai, Shenghua Gao, and Bo Hu. Locality-constrained spatial transformer network for video crowd counting. In *ICME*, pages 814–819. IEEE, 2019. 1, 2

[16] Junyu Gao, Tao Han, Yuan Yuan, and Qi Wang. Learning independent instance maps for crowd localization. *arXiv preprint arXiv:2012.04164*, 2020. 2

[17] Junyu Gao, Tao Han, Yuan Yuan, and Qi Wang. Domain-adaptive crowd counting via high-quality image translation and density reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 4

[18] Junyu Gao, Yuan Yuan, and Qi Wang. Feature-aware adaptation and density alignment for crowd counting in video surveillance. *IEEE transactions on cybernetics*, 51(10):4822–4833, 2020. 1

[19] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, pages 2547–2554, 2013. 2

[20] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *ECCV*, pages 532–546, 2018. 2, 4

[21] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, pages 784–799, 2018. 4

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, pages 1446–1453. IEEE, 2009. 3

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25:1097–1105, 2012. 6

[25] Haopeng Li, Lingbo Liu, Kunlin Yang, Shinan Liu, Junyu Gao, Bin Zhao, Rui Zhang, and Jun Hou. Video crowd localization with multi-focus gaussian neighbor attention and a large-scale benchmark. *arXiv preprint arXiv:2107.08645*, 2021. 6

[26] Wei Li, Hongliang Li, Qingbo Wu, Xiaoyu Chen, and King Ngi Ngan. Simultaneously detecting and counting dense vehicles from drone images. *IEEE TIE*, 66(12):9651–9662, 2019. 6

[27] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018. 1, 2, 6

[28] Ningxin Liang, Guile Wu, Wenxiong Kang, Zhiyong Wang, and David Dagan Feng. Real-time long-term tracking with prediction-detection-correction. *IEEE Transactions on Multimedia*, 20(9):2289–2302, 2018. 3

[29] Cao Lijun and Huang Kaiqi. Video-based crowd density estimation and prediction system for wide-area surveillance. *China Communications*, 10(5):79–88, 2013. 2

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 4

[31] Bo Liu and Nuno Vasconcelos. Bayesian model adaptation for crowd counts. In *ICCV*, pages 4175–4183, 2015. 2

[32] Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In *CVPR*, pages 1217–1226, 2019. 2, 4

[33] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, and Liang Lin. Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4823–4833, 2021. 2

[34] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1774–1783, 2019. 2

[35] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Estimating people flows to better count them in crowded scenes. In *ECCV*, pages 723–740. Springer, 2020. 2

[36] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In *CVPR*, pages 6469–6478, 2019. 2

[37] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, page 103448, 2020. 3

[38] Zheng Ma and Antoni B Chan. Counting people crossing a line using integer programming and local features. *IEEE TCSVT*, 26(10):1955–1969, 2015. 1, 3

[39] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *ICCV*, pages 6142–6151, 2019. 1, 2

[40] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957. 5, 7, 8

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037, 2019. 6

[42] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 5

[43] Weihong Ren, Xinchao Wang, Jiandong Tian, Yandong Tang, and Antoni B Chan. Tracking-by-counting: Using network flows on crowd density maps for tracking multiple targets. *IEEE TIP*, 30:1439–1452, 2020. 3

[44] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *CVPR*, pages 4031–4039, 2017. 6

[45] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020. 5, 6

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[47] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 5

[48] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *ICCV*, pages 3365–3374, 2021. 2, 4

[49] Shijie Sun, Naveed Akhtar, Huansheng Song, Chaoyang Zhang, Jianxin Li, and Ajmal Mian. Benchmark data and method for real-time people counting in cluttered scenes using depth sensors. *IEEE TITS*, 20(10):3599–3612, 2019. 2, 3

[50] Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. Tracking pedestrian heads in dense crowd. In *CVPR*, pages 3865–3875, 2021. 2, 3, 6, 7, 8

[51] UN. World population prospects 2019, 2019. 1

[52] Xuan-Thuy Vo. Phdtt results. [Online]. https://motchallenge.net/method/HT=7&chl=21. 6, 7

[53] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *CVPR*, pages 1974–1983, 2021. 2, 4

[54] Jia Wan, Wenhan Luo, Baoyuan Wu, Antoni B Chan, and Wei Liu. Residual regression with semantic prior for crowd counting. In *CVPR*, pages 4036–4045, 2019. 1

[55] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. *arXiv preprint arXiv:2009.13077*, 2020. 2

[56] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE TPAMI*, 2020. 2

[57] Qi Wang, Tao Han, Junyu Gao, and Yuan Yuan. Neuron linear transformation: Modeling the domain shift for crowd counting. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2

[58] Xingjiao Wu, Baohan Xu, Yingbin Zheng, Hao Ye, Jing Yang, and Liang He. Fast video crowd counting with a temporal aware network. *Neurocomputing*, 403:13–20, 2020. 1, 2

[59] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *ICCV*, pages 5151–5159, 2017. 1, 2

[60] Qi Zhang, Wei Lin, and Antoni B Chan. Cross-view cross-scene multi-view crowd counting. In *CVPR*, pages 557–567, 2021. 2

[61] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, pages 1–19, 2021. 2, 6, 7

[62] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016. 2, 6

[63] Zhuoyi Zhao, Hongsheng Li, Rui Zhao, and Xiaogang Wang. Crossing-line crowd counting with two-phase deep neural networks. In *ECCV*, pages 712–726. Springer, 2016. 1, 3, 7

[64] 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 TCSVT*, 29(3):787–799, 2018. 1, 3