Weakly Supervised Video Individual Counting

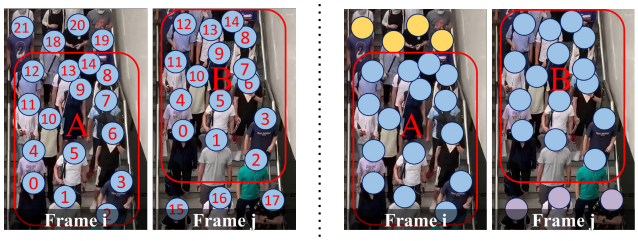
Xinyan Liu[1] Guorong Li[1,3*] Yuankai Qi[2] Ziheng Yan[1]
Zhenjun Han[1] Anton van den Hengel[5] Ming-Hsuan Yang[6] Qingming Huang[1,3,4]
[1] University of Chinese Academy of Science; [2] Macquarie University;
[3] Key Lab of Big Data Mining and Knowledge Management, UCAS, China;
[4] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS;
[5] Australian Institute for Machine Learning, The University of Adelaide;
[6] University of California at Merced.

Abstract

Video Individual Counting (VIC) aims to predict the number of unique individuals in a single video. Existing methods learn representations based on trajectory labels for individuals, which are annotation-expensive. To provide a more realistic reflection of the underlying practical challenge, we introduce a weakly supervised VIC task, wherein trajectory labels are not provided. Instead, two types of labels are provided to indicate traffic entering the field of view (inflow) and leaving the field view (outflow). We also propose the first solution as a baseline that formulates the task as a weakly supervised contrastive learning problem under group-level matching. In doing so, we devise an end-to-end trainable soft contrastive loss to drive the network to distinguish inflow, outflow, and the remaining. To facilitate future study in this direction, we generate annotations from the existing VIC datasets SenseCrowd and CroHD and also build a new dataset, UAVVIC. Extensive results show that our baseline weakly supervised method outperforms supervised methods, and thus, little information is lost in the transition to the more practically relevant weakly supervised task. The code and trained model can be found at CGNet.

1. Introduction

Video Crowd Counting (VCC) has garnered much interest due to its broad range of practical applications, particularly in crowd safety management. This task requires a model to count the number of people in each frame of a video. A limitation of VCC [12, 51, 53] is that it gives an imprecise estimate of the number of unique individuals appearing in a video sequence, as people are counted multiple times if they appear in several frames. To overcome this drawback, Video Individual Counting (VIC) was introduced, wherein a method must count the total number of people with unique identities appearing in a video sequence.



(a) Annotation of VIC method (b) Annotation of our method

Figure 1. Existing VIC method requires a unique label indicating the position of each human in each frame, and these labels are consistent across frames. Weakly supervised VIC only requires labels indicating each human position and whether they are an inflow/outflow pedestrian (purple/yellow in (b)) in the current frame. The transition from individual-level labels to identity-agnostic group-level labels represents a significant reduction in the labeling effort.

The apparent approach to crowd counting is to count the people who appear in the first frame and add the number of people who come into the camera's field of view in later frames (the inflow). Following this principle, Han *et al*. [14] devised DRNet, the only existing published method for VIC, identifying repeated observations of individuals across consecutive frames based on their appearance, simultaneously predicting inflow and outflow. Although the number of outflows does not contribute to the final count, DRNet finds it helps determine individual associations. Han *et al*. [14] also explored using conventional multi-object track-

*Corresponding author.

This work was supported in part by the National Natural Science Foundation of China under Grants 62272438, 62236008, U21B2038, and 61931008, in part by the Key Deployment Program of the Chinese Academy of Sciences under Grant KGFZD145-23-18, and in part by the Fundamental Research Funds for Central Universities (E2ET1104). Yuankai Qi, Anton van den Hengel, and Ming-Hsuan Yang are not supported by the aforementioned fundings.

ing (MOT) methods [7, 41, 44, 54, 56] to tackle the VIC problem. Unfortunately, it turns out MOT methods suffer from poor accuracy. More notably, DRNet and MOT-based methods require trajectory labels (or similar individual association labels) to supervise identity association, which is highly annotation-expensive.

Our key insight is that counting people in the previous and current frames does not require accurate identity associations for those appearing in both frames. For example, as long as we can predict individual 15∼17 as entering in Fig. 1a, even if we arbitrarily associate individuals 0∼14 in frame i to individuals 0∼14 in frame i, we can still correctly infer the crowd counts. *We can still count crowds if we relax individual-level identity associations to group-level associations and do not require trajectory annotations.* Thus, we just need to annotate inflow and outflow people (then we can derive the individual exists in two frames), which reduces annotation costs compared to creating individual pairwise target associations for each observed pedestrian between neighboring frames. We name crowd counting with such annotations as Weakly supervised Video Individual Counting (WVIC).

To address the WVIC task, we propose a benchmark method based on Contrastive learning with Group-level matching, namely CGNet. We design a soft contrastive loss to drive the network to learn discriminative representations that can facilitate identifying the required group associations and, thus, the inflow. Moreover, to better represent each individual, we design a memory-based individual count predictor, where historical templates of individuals are stored and updated in memory to enhance the robustness of association during inference. Considering that the two existing datasets for VIC are all captured by static cameras, we collect a weakly supervised VIC dataset based on moving UAVs, named UAVVIC. This dataset provides about 400,000 inflow/outflow and bounding box labels for four categories: pedestrian, car, bus, and van. Although devised for crowd counting, our proposed baseline method also performs well in counting other objects.

Our main contributions are as follows:

- We propose WVIC, a weakly supervised video individual counting task. This task does not require expensive per-target trajectory annotations and only requires two types of identity-agnostic annotations.
- We automatically reannotate two existing datasets, CroHD and SenseCrowd, and collect a new dataset, UAVVIC, to pave the way for future studies.
- We propose a strong baseline, CGNet, equipped with a newly designed group level matching soft contrastive loss, performing favorably against the supervised methods on the three datasets mentioned above.

2. Related Works

Video Crowd Counting (VCC) estimates the number of people in each video frame. Most of the existing VCC methods [2, 5, 6, 11, 12, 30, 51, 53, 58] can be divided into two categories according to their solved problems: region of interest (ROI) and line of interest (LOI). ROI methods [11, 12, 51, 53] detect pedestrians within a specific region (or the whole image), and they focus on leveraging temporal context information to improve the prediction of the current frame. LSTN [12] models the group flow of crowds in local regions. Xiong *et al*. in [53] fuse history frame features using a ConvLSTM. TAN [51] explores context information from adjacent density maps. CLRNet proposes a local self-attention module to help the model focus on highly related regions in adjacent frames. ROI methods can not be used in the VIC task. LOI methods [2, 5, 6, 30, 57, 58] counts people passing through a specific line. Most of the existing LOI methods [2, 5, 6] apply blobs to crowds, counting objects when the whole blobs cross the line. Ma *et al*. in [30] sample fixed line width areas within temporal image slices and accumulated crowds in this area. Zheng *et al*. [58] instead sum the number of people within neighboring blocks near a line based on local velocity. Zhao *et al*. in [57] train a crowd velocity map predictor with the help of trajectory labels. In practice, LOI methods perform poorly on the VIC task as people can enter or leave the field of view in any direction, making it difficult to find a suitable virtual line.

Multiple Object Tracking (MOT) aims to predict the trajectories for multiple targets in a video. Object association is a main challenge in MOT, establishing correspondences between objects across frames to generate targets' trajectories. Existing object association methods in MOT include Probabilistic Data Association (PDA) [36, 37], Joint Probabilistic Data Association (JPDA) [16, 23, 29, 40], and graph matching methods like the Hungarian algorithm [3, 50]. Recent research has started to use more powerful feature representations to enhance the robustness of data association. Deep learning-based methods like LSTM can model temporal information to handle long-term occlusion [20, 34, 45]. Many works [4, 15, 46] also focus on the association of tracklets. Specifically, they associate multiple tracklets over longer periods using a graph model to combine the final trajectory, which can relieve the effects of lost tracked targets. These methods perform poorly on the VIC task, however, due to the inevitable ID switching problem [14]. Additionally, they require trajectory labels to train their models.

Video Individual Counting (VIC). To the best of our knowledge, DRNet [14] is the only method for the VIC task. It predicts the number of initial pedestrians and matches pedestrians in adjacent frames by adapting a differentiable optimal transport loss. Although the VIC task only requires the total count in the video, DRNet needs individual

matching between frames. WVIC avoids the requirement for such matching annotation, which enables easy expansion of dataset scales and thus facilitates future large deep learning models.

3. A Benchmark Method

The goal of Weakly-supervised Video Individual Counting (WVIC) is to predict the number of unique individuals ($\hat{N}$) in a video provided a series of its frames with δ as the sampling rate. Given a video $\mathbf{V}$ consisting of T frames $\mathbf{V}^1, \cdots, \mathbf{V}^T$, the label information comprises coordinate/inflow/outflow data for each individual. Specifically, letting $\mathbf{P}^i = [p_1^i, ..., p_{n_i}^i]$ denote the annotation of the coordinates of the center of the people in frame $\mathbf{V}^i$, we sample the frame pairs $\{(\mathbf{V}^1, \mathbf{V}^{\delta+1}), (\mathbf{V}^{\delta+1}, \mathbf{V}^{2\delta+1}), ..., (\mathbf{V}^{T-\delta}, \mathbf{V}^T)\}$, where δ is an integer frame offset. The weakly supervised inflow label $\mathbf{I}^i = [\mathbf{I}_1^i, ..., \mathbf{I}_{n_i}^i]$ and outflow label $\mathbf{O}^i = [\mathbf{O}_1^i, ..., \mathbf{O}_{n_i}^i]$. If the u-th person in frame i doesn't appear in $\mathbf{V}^{i-\delta}$, then $\mathbf{I}_u^i = 1$; otherwise, $\mathbf{I}_u^i = 0$. For $\mathbf{O}^i$, if the u-th person in frame V^i doesn't appear in $\mathbf{V}^{i+\delta}$, then $\mathbf{O}_u^i = 1$; otherwise $\mathbf{O}_u^i = 0$. If, by a slight abuse of notation, we refer to people by their coordinates in particular images, we see that the set $\{p_u^i | \mathbf{O}_u^i = 0\}$ and $\{p_u^{i+\delta} | \mathbf{I}_u^i = 0\}$ represent the people that appear both in $\mathbf{V}^i$ and $\mathbf{V}^{i+\delta}$.

3.1. Weakly Supervised Representation Learning

As shown in Fig. 2, the inference pipeline of our CGNet consists of three components: an image-level locator to generate coordinates for pedestrians, an encoder to produce representations based on pedestrians' coordinates, and a Memory-based individual Count Predictor (MCP) to predict the final inflow count based on the representations. The locator is trained independently with the coordinate annotations, and existing image crowd localization networks such as FIDT [25] can be used. The encoder is trained with our Weakly Supervised Representation Learning method (WSRL) to extract the discriminative features of each individual. It can be any feature extractor, such as ConvNext [26]. MCP does not need training.

Provided two δ-adjacent frames $\mathbf{V}^i$ and $\mathbf{V}^j$ and the corresponding annotations $\mathbf{P}^i$, $\mathbf{P}^j$, $\mathbf{O}^i$ and $\mathbf{I}^j$, we group the individuals into four sets: $\mathbf{X} = \{p_u^i | \mathbf{O}_u^i = 0\}$, $\mathbf{Y} = \{p_u^j | \mathbf{I}_u^j = 0\}$, denoting shared individuals in the previous frame and the current frame, respectively; $\hat{\mathbf{X}} = \mathbf{P}^i - \mathbf{X}$, and $\hat{\mathbf{Y}} = \mathbf{P}^j - \mathbf{Y}$, denoting outflow individuals and inflow individuals, respectively. An encoder learns representations $\mathbf{F}$ for each pedestrian based on the crops around those coordinates. Specifically, to generate the u-th feature f_u^i in $\mathbf{F}$, if the locator outputs points, we extract a rectangle patch of size 96×64 centered at the predicted coordinate p_u^i, and if the outputs are bounding boxes, we directly crop the bounding box from the original picture. Each cropped patch is resized to 224×224 and fed into the encoder, a ConvNext-S [26], generating a feature of $7 \times 7 \times 768$, which is flattened and normalized into a 1D vector f_u^i to represent the final feature for the individual:

$$\begin{aligned} \mathbf{F}_{\mathbf{X}}^i = \text{Encoder}(\mathbf{X}), \mathbf{F}_{\hat{\mathbf{X}}}^i &= \text{Encoder}(\hat{\mathbf{X}}) \\ \mathbf{F}_{\mathbf{Y}}^j = \text{Encoder}(\mathbf{Y}), \mathbf{F}_{\hat{\mathbf{Y}}}^j &= \text{Encoder}(\hat{\mathbf{Y}}) \end{aligned} \tag{1}$$

Then, a similarity matrix $\mathbf{S} \in \mathbb{R}^{n_i \times n_j}$ between $\mathbf{F}^i = \left[\mathbf{F}_{\mathbf{X}}^i, \mathbf{F}_{\hat{\mathbf{X}}}^i\right]$ and $\mathbf{F}^j = \left[\mathbf{F}_{\mathbf{Y}}^j, \mathbf{F}_{\hat{\mathbf{Y}}}^j\right]$ is computed by:

$$\mathbf{S} = \begin{bmatrix} (\mathbf{F}_{\mathbf{X}}^i)^\top \mathbf{F}_{\mathbf{Y}}^j & (\mathbf{F}_{\mathbf{X}}^i)^\top \mathbf{F}_{\hat{\mathbf{Y}}}^j \\ (\mathbf{F}_{\hat{\mathbf{X}}}^i)^\top \mathbf{F}_{\mathbf{Y}}^j & (\mathbf{F}_{\hat{\mathbf{X}}}^i)^\top \mathbf{F}_{\hat{\mathbf{Y}}}^j \end{bmatrix} = \begin{bmatrix} \mathbf{S}_0 & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{S}_3 \end{bmatrix}. \tag{2}$$

The matrix $\mathbf{S}$ is divided into four parts: $\mathbf{S}_0 \in \mathbb{R}^{m_{i,j} \times m_{i,j}}$, $\mathbf{S}_1 \in \mathbb{R}^{m_{i,j} \times (n_j - m_{i,j})}$, $\mathbf{S}_2 \in \mathbb{R}^{(n_i - m_{i,j}) \times m_{i,j}}$, $\mathbf{S}_3 \in \mathbb{R}^{(n_i - m_{i,j}) \times (n_j - m_{i,j})}$, where $m_{i,j} = (\mathbf{1} - \mathbf{I}^{\mathbf{j}})^T (\mathbf{1} - \mathbf{O}^i)$ denotes the number of shared individuals in $\mathbf{V}^i$ and $\mathbf{V}^j$. $\mathbf{S}_0$ is the similarity matrix of $\mathbf{X}$ and $\mathbf{Y}$, while $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$ are the similarity matrix of $\mathbf{X}$ and $\hat{\mathbf{Y}}$, of $\hat{\mathbf{X}}$ and $\mathbf{Y}$, of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$, respectively.

To pull the matched groups ($\mathbf{X}$ and $\mathbf{Y}$) closer and push away individual pairs from unmatched groups, we propose a weakly supervised Group-level Matching Loss (**GML**) to constrain different parts of $\mathbf{S}$.

Constraint for $\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2$. $\mathbf{S}_0$ denotes the similarity among individuals shared between two frames, and it does not matter whether the individuals are correctly matched as long as we can assign them a one-to-one match. As for $\mathbf{S}_1$ and $\mathbf{S}_2$, they are the similarities of different identities and thus should be zero.

To this end, we introduce a latent variable matrix $\Omega \in \mathbb{R}^{m_{i,j} \times m_{i,j}}$, where the (u, v)-th element $\Omega_{u,v}$ denote the probability that the u-th pedestrian in $\mathbf{X}$ is matched with v-th pedestrian in $\mathbf{Y}$. To increase the similarity between the pairs that have a higher matching probability and decrease the similarity between the pairs that have a lower matching probability, inspired by [32, 39], we define the soft contrastive loss as:

$$\begin{aligned} \mathcal{L}_{scon}(i) = \min_{\Omega} - \sum_{u=1}^{m_{i,j}} \sum_{v=1}^{m_{i,j}} \Omega_{u,v} C_{u,v}, \\ s.t. \quad \mathbf{1}_{m_{i,j}}^T \Omega = \mathbf{1}, \quad \Omega \mathbf{1}_{m_{i,j}} = \mathbf{1}, 0 \leq \Omega_{u,v} \leq 1, \end{aligned} \tag{3}$$

where $C_{u,v}$ denotes the contrastive similarity [10], and is calculated as,

$$C_{u,v} = \frac{e^{\frac{1}{\gamma}\mathbf{S}_{u,v}}}{e^{\frac{1}{\gamma}\mathbf{S}_{u,v}} + \sum\limits_{1 \leq u' \leq n_i}^{u' \neq u} e^{\frac{1}{\gamma}\mathbf{S}_{u',v}} + \sum\limits_{1 \leq v' \leq n_j}^{v' \neq v} e^{\frac{1}{\gamma}\mathbf{S}_{u,v'}}}, \tag{4}$$

where γ is a temperature hyper-parameter [52] and $\mathbf{S}_{u,v} = \langle f_u^i, f_v^j \rangle$ is the (u, v)-th element of $\mathbf{S}$, calculated by the u-th feature in $\mathbf{F}^i$ and the v-th feature in $\mathbf{F}^j$. As $\Omega_{u,v}$ describes the matching probability, our soft contrastive loss

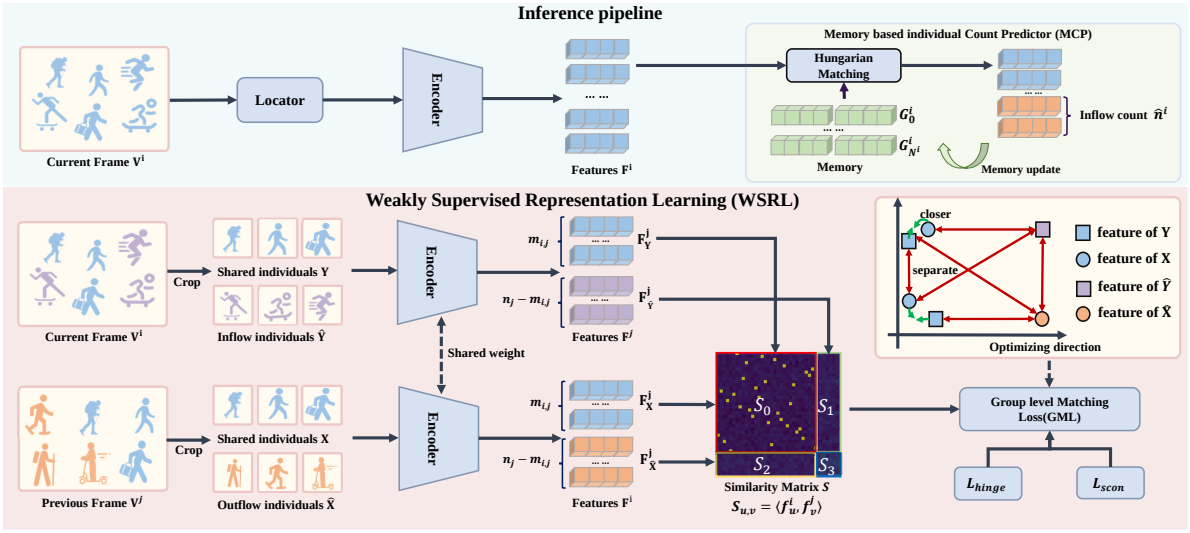


Figure 2. The inference pipeline of our CGNet and the weakly supervised representation learning method (WSRL). The pipeline comprises a frame-level crowd locator, an encoder, and a Memory-based individual count predictor (MCP). The locator predicts the coordinates for pedestrians. The encoder generates representations for each individual, and MCP predicts inflow counts and updates the individual templates stored in the memory. To pull the matched groups ($\mathbf{X}$ and $\mathbf{Y}$) closer and push away individual pairs from unmatched groups, WSRL exploits inflow and outflow labels to optimize the encoder with a novel Group level Matching Loss (GML), which consists of a soft contrastive loss ($\mathcal{L}_{scon}$) and a hinge loss ($\mathcal{L}_{hinge}$).

evaluates the expectation of contrastive loss. It should be noting that in $C_{u,v}$, (f_u^i, f_v^j) is considered as positive pairs, while $\{(f_u^i, f_{v'}^j)|v' \neq v\}$ and $\{(f_{u'}^i, f_v^j)|u' \neq u\}$ are considered as negative pairs. Thus, $\mathcal{L}_{scon}$ can encourage $\mathbf{S}_0$ to be a permutation matrix, which indicates that there exists a one-to-one match between $\mathbf{X}$ and $\hat{\mathbf{Y}}$, and drive $\mathbf{S}_1, \mathbf{S}_2$ to be zero matrices.

The problem defined by Eq. (3) is a typical balanced Optimal Transport (OT [35]) problem which transports $\mathbf{1}_{m_{i,j}}$ to $\mathbf{1}_{m_{i,j}}$ using $\mathbf{1} - C$ as the cost matrix and Ω is the transport matrix. Therefore, we consider $\mathcal{L}_{scon}(i)$ as an OT loss, which can be solved with the Sinkhorn algorithm [38].

Constraint for $\mathbf{S}_3$. Obviously, the elements in $\mathbf{S}_3$ measure the similarity of different individuals, so they should be as small as possible. However, directly constraining $\mathbf{S}_3$ to zero matrix will make features from $\hat{\mathbf{X}}, \hat{\mathbf{Y}}$ easily collapse to zero-vector. To this end, the widely used Hinge L1 loss [18] is adopted to constrain $\mathbf{S}_3$:

$$\mathcal{L}_{hinge}(i) = \frac{\sum Relu(\mathbf{S_3} - \theta)}{(n_i - m_{i,j})(n_j - m_{i,j})}, \tag{5}$$

where θ is a threshold used to ignore small values in $\mathbf{S}_3$.

Finally, the group-level matching loss of the training set $\mathbf{V}$ is formulated as,

$$\mathbf{GML}(\mathbf{V}) = \sum_{0<i<T, i=k\delta+1} \left(\mathcal{L}_{scon}(i) + \mathcal{L}_{hinge}(i)\right). \tag{6}$$

3.2. Memory-based Count Predictor

To handle individuals' appearance variance and re-entering cases, inspired by [22, 33, 55], we propose a memory-based count predictor (MCP) to reason the final count for the video, which stores templates of recently appeared individuals in a flexible memory. Specifically, when processing frame $\mathbf{V}^{i+\delta}$, the memory is denoted as $\mathcal{G}^i = \{\mathcal{G}_1^i, \mathcal{G}_2^i, ..., \mathcal{G}_{\mathbf{N}^i}^i\}$, where $\mathbf{N}^i$ is the memory size at time i. $\mathcal{G}_u^i = \{g_{u,0}^i, ..., g_{u,k}^i\}$ is a set of templates for u-th pedestrian in the memory, and k is the number of templates stored for u-th pedestrian.

We first use the crowd locator to predict the coordinates of pedestrians in $\mathbf{V}^{i+\delta}$, and then extract their corresponding features, denoted as $\{f_1^{i+\delta}, f_2^{i+\delta}, \cdots, f_{n_{i+\delta}}^{i+\delta}\}$. To associate the individuals in $\mathbf{V}^{i+\delta}$ with templates in memory, we first define the cost $\hat{C}_{u,u'}$ of matching the u-th pedestrian with the u'-th template in $\mathcal{G}^i$ as

$$\hat{C}_{u,u'} = \max_v (1 - \langle f_u^{i+\delta}, g_{u',v}^i \rangle). \tag{7}$$

Then, we generate an optimal one-to-one match $\pi \in \{0,1\}^{n_{i+\delta} \times \mathbf{N}^i}$ (assuming $\mathbf{N}^i \geq n_i$) by solving the following problem with Hungarian Algorithm [21]:

$$\begin{aligned} \pi &= \arg\min_{\pi} \sum_{u,u'} \pi_{u,u'} \hat{C}_{u,u'} \\ s.t. \quad &\forall u', \sum_u \pi_{u,u'} = 1, \quad \forall u, \sum_u \pi_{u,u'} \leq 1. \end{aligned} \tag{8}$$

If the matching cost for u meets $\sum_{u'=1}^{\mathbf{N}^i} \pi_{u,u'} \hat{C}_{u,u'} > \zeta$,

u will be judged as an inflow pedestrian. Otherwise, if $\pi_{u,u'} = 1$, u is associated with $g^i_{u'}$.

Template update. Similar to [22], $\mathcal{G}^i_u$ has a time-to-live factor (**ttl**). $\mathbf{ttl}_u$ will decrease one if $\mathcal{G}^i_u$ has no mapping to any pedestrian feature in $\mathbf{F}^{i+\delta}$. If $\mathbf{ttl}_u$ decreases to zero, the template $\mathcal{G}^i_u$ will be dropped. Meanwhile, whenever $\mathcal{G}^i_u$ is associated with a pedestrian in the current frame, $\mathbf{ttl}_u$ will be reset to **ttlmax**. This time-to-live threshold allows us to save some recently appeared targets while discarding those that have not appeared for a long time. If pedestrian u is associated with u'-th template, we add $f^{i+\delta}_u$ to $\mathcal{G}^i_{u'}$ to update the template

$$\mathcal{G}^{i+\delta}_{u'} \leftarrow \mathcal{G}^i_{u'} \cup \{f^{i+\delta}_u | \pi_{u,u'} = 1, \sum_{u'=1}^{\mathbf{N}^i} \pi_{u,u'} \hat{C}_{u,u'} \leq \zeta\}. \quad (9)$$

If pedestrian u is an inflow pedestrian, $\{f^{i+\delta}_u\}$ will be added to the memory $\mathcal{G}^{i+\delta}$ as a new template.

The output number of the total individuals in video I is then calculated via

$$\hat{N} = n_1 + \sum_{\delta+1 \leq i \leq T, i=k\delta+1} \hat{n}_i, \quad (10)$$

where $\hat{n}_i$ is the number of inflow pedestrians at time i.

4. Experiments

4.1. Dataset

We test our CGNet on three datasets: CroHD [42], SenseCrowd [24], and UAVVIC. CroHD has four training videos and five testing videos. SenseCrowd contains 634 videos, and we split the train, validation, and test dataset following DRNet [14]. UAVVIC is our proposed dataset collected by a moving UAV camera in various scenes, including campus, beach, car park, highway, city road, and square. UAVVIC consists of 221 videos (100 for training, 100 for testing, and 21 for validation), and 5,396 frames are sampled with 3s as the interval. Annotations consist of 398,158 bounding boxes, and group-level matching labels in neighbor frames are provided. The resolutions of UAVVIC are 4K and 1080P for better capture of tiny pedestrians from drone view. A detailed comparison of UAVVIC with existing video crowd counting datasets [8, 9, 13, 24, 42] is shown in Tab. 1. UAVVIC provided a moving shot with a larger range of pedestrians. More details about UAVVIC are provided in the supplementary materials.

4.2. Metrics

Mean Absolute Error (MAE), Mean Square Error (MSE), and Weighted Relative Absolute Errors (WRAE) are used for evaluation. The first two metrics are common metrics applied in VCC. However, unlike VCC schemes, we only count the same person once in the video. WRAE [14] is proposed to balance the performance on videos with different lengths and pedestrian numbers: $WRAE = \sum_{i=1}^{K} \frac{T_i}{\sum_{j=1}^{K} T_j} \frac{|N_i - \hat{N}_i|}{N_i}$, where K is the total number of videos, and T_i is the length of video i, $N_i/\hat{N}_i$ denotes the number of ground truth/predicted counts in i-th video.

Name	Resolution	Range	Moving shot	Point	Box	Trajectory	In-Out
Mall [9]	640*480	13-53	✗	✓	✗	✗	✗
UCSD [8]	238*158	11-46	✗	✓	✗	✗	✗
FDST [13]	1920*1080 1280*720	9-57	✗	✓	✗	✗	✗
SenseCrowd [24]	-	1-296	✗	✓	✓	✓	✗
CroHD [42]	-	25-346	✗	✓	✓	✓	✗
UAVVIC	3840*2160 1920*1080	0-735	✓	✓	✓	✗	✓

Table 1. Comparison of different video crowd counting datasets.

4.3. Implementation Details

Training. If there are no special instructions, the locator we applied is a well-known crowd localization model, FIDT [25]. All parameters are official from FIDT, except that we replaced the backbone from HRNet-W48 [47] to HRNet-W18 [47] for efficient inference. The learning rate is set as $1e^{-4}$ along with AdamW [28] as the optimizer and applying the pre-train weight from Timm [49]. All related models can be trained on one RTX3090 (24G memory).

Testing. In the testing phase, the time interval δ is set as $3s$ following the setting in [14], the threshold ζ is set as 0.7, γ in Eq. (4) is set as 10, and the max time-to-live (**ttlmax**) is set as 3, the max size of memory (**memmax**) is set as 5.

4.4. Overall Performance

Comparison Methods. To the best of our knowledge, only DRNet[14] has been specifically designed for the VIC task. To evaluate the effectiveness of our proposed CGNet, we also tested two categories of approaches that could be applied to VIC: 1) Multi-object tracking (MOT) methods, including HeadHunter-T [43], FairMOT [56], PHDTT [44], SMILE [48], SparseTrack [27], Deep-OC-SORT [31], and BoT-SORT [1], where we consider the total number of indices in the entire video as individual count. We set the frame rate to 1 FPS for HeadHunter-T and 0.33 FPS for the other MOT methods to obtain better VIC performance. 2) A recent cross-line video crowd-counting, LOI [57], which also requires trajectory labels in training.

Results on SenseCrowd: As shown in Tab. 2, it is evident that our CGNet exhibits superior performance on this larger dataset, SenseCrowd. Particularly, CGNet improved the overall MAE by about 28% compared to DRNet. Furthermore, we achieve the lowest MAE across all density levels except for D0. Compared to the existing lowest MAE on D1, D2, D3, and D4, the performance improvements of our CGNet are about 27%, 63%, 37%, and 17%, respectively. In Fig. 3, we show qualitative results on three representative scenes,i.e., mall, station, and outdoor plaza. Although the inflows in the current flow are not concentrated, CGNet

(a) Previous frame (0s)

GT inflow: 4 Predict inflow: 4 GT inflow: 11 Predict inflow: 10 GT inflow: 15 Predict inflow: 14

(b) Current frame (3s)

Figure 3. Results on the SenseCrowd dataset. Red dots are predictions of the locator. Blue/green/yellow circles are correct predicted inflow, error predicted inflow, and missed inflow, respectively.

Method	MAE	MSE	WRAE(%)	MAE on five different density levels				
				D0	D1	D2	D3	D4
FairMOT [56]	35.4	62.3	48.9	13.5	22.4	67.9	84.4	145.8
HeadHunter-T [43]	30.0	50.6	38.6	11.8	25.7	56.0	92.6	131.4
SMILE [48]	27.22	36.7	32.5	9.2	21.0	33.8	76.5	203.5
SparseTrack [27]	28.36	43.8	30.8	9.9	21.5	42.1	89.2	189.5
Deep-OC-SORT [31]	26.04	43.6	29.5	9.5	25.1	29.1	60.3	156.5
BoT-SORT [1]	24.67	32.4	28.3	8.7	20.5	31.1	70.2	165.1
LOI [57]	24.7	33.1	37.4	12.5	25.4	39.3	39.6	86.7
DRNet [14]	12.3	24.7	12.7	4.1	8.0	23.3	50.0	77.0
CGNet (Ours)	8.86	17.69	12.6	5.0	5.8	8.5	25.0	63.4

Table 2. Performance comparison on SenseCrowd. The density labels D0-D4 are defined in [14], where D0, D1, D2, D3, and D4 correspond to the count range [0, 50), [50, 100), [100, 150), [150, 200) and [200, $+\infty$), respectively. The best values are highlighted in blue font.

Method	MAE	MSE	WRAE(%)	MAE on five testing scenes				
				CroHD11	CroHD12	CroHD13	CroHD14	CroHD15
PHDTT [44]	2130.4	2808.3	401.6	247	3793	4794	491	1327
FairMOT [56]	256.2	300.8	44.1	11	427	284	408	1000
HeadHunter-T [43]	253.2	351.7	32.7	65	101	515	582	1003
SMILE [48]	257.6	334.5	40.5	16	231	156	649	246
SparseTrack [27]	176.6	208.6	27.6	29	172	258	336	88
Deep-OC-SORT [31]	165.2	195.9	33.1	68	351	186	161	60
BoT-SORT [1]	154.8	176.42	27.4	49	210	138	286	91
LOI [57]	305.0	371.1	46.0	60	243	458	630	131
DRNet [14]	141.1	192.3	27.4	31	338	18	255	61
CGNet (Ours)	75.0	95.1	14.5	7	72	14	144	138

Table 3. Performance comparison on CroHD dataset. CroHD11-CroHD15 are five test scenes labeled in [14]. The best values are highlighted in blue font.

can accurately predict them. More visualization results can be found in supplementary material.

Results on CroHD. Comparison with existing methods on CroHD is shown in Tab. 3. Our CGNet achieved the lowest MAE, MSE, and WRAE. Especially, except for CroHD15, our method obtained the lowest MAE across the other four testing scenes. As CGNet is trained with only point annotations, it cannot crop the patch of optimal size to extract features on CroHD15. It should be noted that on CroHD13 and CroHD14, which have the highest average density (245.9 and 259.6 person/frame), the performances of all the compared methods except for DRNet degrade sig-

Method	MAE/MSE/WRAE	MAE/MSE/WRAE on four different categories			
		Pedestrian	Car	Bus	Van
BoT-SORT [1]	49.9/228.2/29.9	92.4/398.2/49.0	59.1/259.1/39.2	23.0/109.6/57.1	25.1/146.2/77.3
Deep-OC-SORT [31]	37.1/115.4/35.2	85.2/213.5/22.1	33.4/113.1/21.0	19.1/ 78.2/26.2	10.7/56.6/32.8
DRNet [14]	18.4/ 41.0/15.2	34.5/ 92.1/19.1	19.0/ 32.7/17.8	10.4/ 22.1/32.2	9.9/17.2/29.4
CGNet (Ours)	12.9/ 37.4/ 12.0	31.2/ 98.4/14.4	16.9/ 45.2/ 7.3	2.1/ 4.1/19.2	1.3/ 2.0/25.1

Table 4. Performance comparison on UAVVIC. All the methods are trained independently with four classes. The values of each entry are MAE/MSE/WRAE. The best values are highlighted in blue font.

Method	MAE	MSE	WRAE(%)	MAE on five different density levels				
				D0	D1	D2	D3	D4
①	13.35	32.3	22.8	9.7	12.3	15.3	26.5	43.4
②	9.84	20.0	13.8	5.3	6.3	12.0	28.8	67.7
③	10.71	17.40	16.9	6.9	8.9	11.7	24.3	50.6
④	8.86	17.69	12.6	5.0	5.8	8.5	25.0	63.4

Table 5. Ablation study on the main components. The density labels D0-D4 are defined in [14], where D0, D1, D2, D3, and D4 correspond to the count range [0, 50), [50, 100), [100, 150), [150, 200) and [200, $+\infty$), respectively. ①, No constrain on $\mathbf{S}_1, \mathbf{S}_2$; ②, w/o Eq. (5); ③, w/o MCP; ④, full CGNet. The best values are highlighted in blue font.

nificantly. In contrast, the performance of our CGNet is relatively stable in these two scenes. Furthermore, even without individual-level matching labels, our CGNet performs favorably against the state-of-the-art approach, DRNet, by a large margin (about 46% on the MAE), demonstrating the effectiveness of our method.

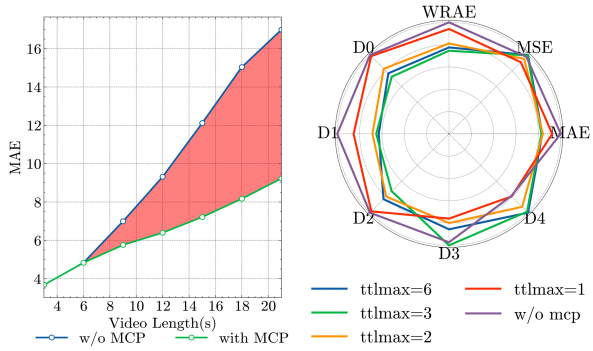
(a) Previous frame (0s) (b) Current frame (3s)

Figure 4. Qualitative visualization of our CGNet on UAVVIC.

Results on UAVVIC. The results on our proposed UAVVIC dataset are shown in Tab. 4. All methods are trained and tested on each category, respectively. Since there is no trajectory annotation on UAVVIC, we train the compared methods [1, 14, 31] on VisDrone [59], a UAV dataset for MOT. Our CGNet shows favorable performance in all categories. The average MAE is improved by about 30% compared to DRNet. In Fig. 4, we show the quality result of our CGNet on the UAVVIC dataset. Although the UAV moves, CGNet can still detect inflow objects accurately.

4.5. Ablation Studies

We conducted several ablation studies on the SenseCrowd. 1). We verify the effectiveness of the main components of our CGNet. 2). We analyze the effect of two main parameters. 3). We report the performance of our CGNet with different locators. 4). We use the trajectories generated by



(a) (b)

Figure 5. Effectiveness of MCP on SenseCrowd. (a) The average MAE with different lengths of videos with and w/o MCP. (b) Performance with different max time-to-live factor **ttlmax** of MCP.

our CGNet as pseudo-labels for existing VIC methods. 5). We compare the time cost of the WVIC and VIC labels.

Effectiveness of Constrain for $\mathbf{S}_1, \mathbf{S}_2$. Eq. (4) constrains $\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2$ together. Without considering un-matched pairs in $\mathbf{X} \times \hat{\mathbf{Y}}$ and $\hat{\mathbf{X}} \times \mathbf{Y}$, it can be changed to constrain only $\mathbf{S}_0$ by remove elements in $\mathbf{S}_1, \mathbf{S}_2$. As shown in Tab. 5, MAE dropped by about 4.51 without constraints on $\mathbf{S}_1$ and $\mathbf{S}_2$ (① VS ④). This is because elements in $\mathbf{S}_1$ and $\mathbf{S}_2$ are related to pedestrians that exist only in one frame. Considering $\mathbf{X} \times \hat{\mathbf{Y}}$ and $\hat{\mathbf{X}} \times \mathbf{Y}$ as negative pairs, our soft contrastive loss can push apart individuals in $\hat{\mathbf{X}}$ ($\hat{\mathbf{Y}}$) with that in $\mathbf{Y}$ ($\mathbf{X}$), thus enhancing the distinguishing ability of the representations.

Effectiveness of Eq. (5). We replace the hinge MAE loss in Eq. (5) to normal L1 loss and derive the method ② in Tab. 5. Compared to L1 loss, our $\mathcal{L}_{hinge}(\cdot)$ gives a soft margin to the inner product of two features other than encouraging them to be 0. Optimizing the inner product of all features toward 0 makes it easier to cause either side to collapse a $\mathbf{0}$ vector, which is not conducive to learning different representations. Compared the performance of ② with that of ④ in Tab. 5, $\mathcal{L}_{hinge}(\cdot)$ improves MAE in all density levels.

Effectiveness of MCP. As shown in Tab. 5, compared ③ with ④, MCP brings about 4.1 improvements on WRAE. Specifically, we evaluate the average MAE of our CGNet on videos with different lengths. As shown in Fig. 5(a), the performance gap between ③ and ④ becomes larger as

δ	MAE	MSE	WRAE(%)	MAE on five different density levels				
				D0	D1	D2	D3	D4
5	10.96	20.26	19.5	6.5	6.9	10.4	29.2	79.9
4	8.71	14.33	12.9	4.2	4.9	10.2	27.8	70.1
3	8.86	17.69	12.6	5.0	5.8	8.5	25.0	63.4
2	9.77	20.01	17.3	7.1	8.8	12.4	17.4	52.6
1	9.56	15.34	16.6	6.8	7.4	10.5	23.1	45.5

Table 6. Ablation study on interval δ. The best values are highlighted in blue font.

Method	Locator	Encoder	MAE	MSE	WRAE
①	FIDT	ConvNext-S	8.86	17.69	9.27
②	Yolov8 nano	ConvNext-S	8.01	16.42	8.85
③	Yolov8 small	ConvNext-S	7.56	15.19	8.05
④	VGG16+FPN	ConvNext-S	9.20	17.45	9.55
⑤	VGG16+FPN	PrRoIPooling	10.01	18.22	9.70
⑥	VGG16+FPN	PrRoIPooling	12.59	23.32	12.58

Table 7. Performance on SenseCrowd of our scheme with different locators and feature extractors. In ⑥, MCP is replaced with the inflow reasoning method in DRNet. The best values are highlighted in blue font.

the length of the video increases. This is because, with the stored templates of individuals, MCP can handle appearance variations and individuals' re-entry, alleviating error accumulation during inflow reasoning.

Effect of ttlmax. The performance with different max long-time-live factor **ttlmax** is shown in Fig. 5(b). Compared with no memory (i.e., **ttlmax** $= 0$), adding memory is helpful for all density levels except for extremely high-density levels (D4). It may be because high-density crowds result in too many candidate individual-level matches, making it difficult for our CGNet to choose the right one. Meanwhile, the performance degrades when **ttlmax** increases from 3 to 6 for that larger **ttlmax** $= 6$ leads to the templates of individuals that appeared a long time ago being stored in the memory. As the probability of these people re-entering the field is low, the recorded templates become distractions.

Effect of Interval δ. The performance with different intervals δ is shown in Tab. 6. The performance on higher-density video is better when δ is smaller. This is because crowds flow faster in those videos, and long intervals may lead to missing pedestrians in sampled frames. The overall best performance is achieved when δ is set as $4s$.

Effect of the Locator. We replaced FIDT in our CGNet with Yolo V8 [19] nano and small and the localization branch (VGG16+FPN) used in DRNet, respectively. Different from FIDT, these three locators are trained with box annotations. For Yolo V8, the predicted bounding boxes are fed to our feature extractor to generate representations of individuals. For a fair comparison with DRNet, we also replace our feature extractor ConvNext-S with PrRoIPooling [17] used in DRNet [14] and remove MCP. As shown in Tab. 7, changing locators can further boost the performance of our method. With the same locator and feature extractor as DRNet (⑤), the MAE of our method is 10.01, 2.99 lower than that of DRNet. Even without MCP (⑥), the MAE of our method is only slightly larger than that of DRNet (12.59 v.s. 12.3), demonstrating that our method can learn effective representations without trajectory labels.

Locater	Ground Truth Labels			Pseudo Labels		
	MAE	MSE	WRAE	MAE	MSE	WRAE
FairMOT	35.4	62.3	48.9	37.2	64.8	49.8
HeadHunter-T	30.0	50.6	38.6	32.5	54.2	39.9
SMILE	27.2	36.7	32.5	27.5	37.0	35.2
SparseTrack	28.4	43.8	30.8	30.5	45.0	31.5
Deep-OC-SORT	26.0	43.6	29.5	29.5	45.3	31.7
BoT-SORT	24.7	32.4	28.3	25.0	35.9	33.8
LOI	24.7	33.1	37.4	26.9	35.2	39.9
DRNet	12.3	24.7	12.7	14.5	25.5	13.3

Table 8. Performance of existing VIC methods trained with different trajectory labels.

$(n_i, n_{i+\delta})$	(9,8)	(60,53)	(116,115)	(289,292)	(645,724)
Association [14]	23.8 s	104.5 s	174.4 s	586.9 s	1663.1 s
In-Out	4.4 s	42.6 s	53.7 s	95.0 s	571.9 s

Table 9. Time cost for a pair of frames with two types of labeling methods. n_i denotes the number of pedestrians in frame i.

Performs as a Pseudo Trajectory Generator. The latent variable Ω is obtained by solving a problem defined by Eq. (3). The Hungarian Algorithm generates a one-to-one match using Ω as the cost matrix. Based on the matching, we generate pseudo trajectory labels, which are used to train existing VIC methods. We show the results on SenseCrowd in Tab. 8. With the generated pseudo label, the performance of all the methods degrades slightly than using the ground truth trajectory labels, demonstrating that exact individual-level matching annotations may not be necessary for VIC.

Label time cost. We compare the time cost of annotating our weakly supervised label and association label [14] on five frame pairs with various densities in Tab. 9. The annotation time for our weakly supervised label is only about 30% of that of the association label, making it easier to build a larger dataset for this VIC.

5. Conclusions

We propose a new task, Weakly supervised Video Individual Counting (WVIC), in this paper. Unlike conventional video individual counting tasks, which require expensive trajectory labels, our WVIC task just requires two types of easily available identity-agnostic annotations. This largely reduces the annotation cost and facilitates future works to upscale data size. Furthermore, we propose a strong baseline for the WVIC task. The key idea is a Group-level Matching Loss (GML), which makes it easier to distinguish inflow individuals from all objects. Last, we have extended two existing datasets and built a totally new dataset for the WVIC task. Our method achieves favorable performance on all three datasets, even compared to supervised methods.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 5, 6, 7

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

[3] Marco Allodi, Alberto Broggi, Domenico Giaquinto, Marco Patander, and Antonio Prioletti. Machine learning in tracking associations with stereo vision and lidar observations for an autonomous vehicle. In *IVS*, pages 648–653, 2016. 2

[4] Seung Hwan Bae and Kuk-Jin Yoon. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *IEEE TPAMI*, 40(3):595–610, 2018. 2

[5] Javier Barandiarán, Berta Murguia, and Fernando Boto. Real-time people counting using multiple lines. In *WIAMIS*, pages 159–162, 2008. 2

[6] Jesús Bescós, José M. Menéndez, and Narciso García. DCT based segmentation applied to a scalable zenithal people counter. In *ICIP*, pages 1005–1008, 2003. 2

[7] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *AVSS*, pages 1–6, 2017. 2

[8] Antoni B. Chan and Nuno Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE TPAMI*, 30(5):909–926, 2008. 5

[9] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *BMVC*, pages 3–14, 2012. 5

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[11] Li Dong, Haijun Zhang, Jianghong Ma, Xiaofei Xu, Yimin Yang, and Q. M. Jonathan Wu. Clrnet: A cross locality relation network for crowd counting in videos. *IEEE TNNLS*, pages 1–15, 2022. 2

[12] 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, 2019. 1, 2

[13] Yanyan Fang, Biyun Zhan, Wandi Cai, Shenghua Gao, and Bo Hu. Locality-constrained spatial transformer network for video crowd counting. *arXiv preprint arXiv:1907.07911*, 2019. 5

[14] Tao Han, Lei Bai, Junyu Gao, Qi Wang, and Wanli Ouyang. DR.VIC: decomposition and reasoning for video individual counting. In *CVPR*, pages 3073–3082, 2022. 1, 2, 5, 6, 7, 8

[15] Yuhang He, Jie Han, Wentao Yu, Xiaopeng Hong, Xing Wei, and Yihong Gong. City-scale multi-camera vehicle tracking by semantic attribute parsing and cross-camera tracklet matching. In *CVPRW*, pages 2456–2465, 2020. 2

[16] Andinet Hunde and Beshah Ayalew. Automated multi-target tracking in public traffic in the presence of data association uncertainty. In *ACC*, pages 300–306, 2018. 2

[17] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tiantian Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, pages 784–799, 2018. 8

[18] Thorsten Joachims. Support vector method for multivariate performance measures. In *ICML*, pages 377–384, 2005. 4

[19] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. https://github.com/ultralytics/ultralytics, 2023. 8

[20] Chanho Kim, Fuxin Li, Mazen Alotaibi, and James M. Rehg. Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking. In *CVPR*, pages 9553–9562, 2021. 2

[21] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 4

[22] L.J. Latecki and R. Miezianko. Object tracking with dynamic template update and occlusion detec. In *ICPR*, volume 1, pages 556–560, 2006. 4, 5

[23] Eui-Hyuk Lee, Qian Zhang, and Taek Lyul Song. Markov chain realization of joint integrated probabilistic data association. *Sensors*, 17(12):2865, 2017. 2

[24] Haopeng Li, Lingbo Liu, Kunlin Yang, Shinan Liu, Junyu Gao, Bin Zhao, Rui Zhang, and Jun Hou. Video crowd localization with multifocus gaussian neighborhood attention and a large-scale benchmark. *IEEE TIP*, 31:6032–6047, 2022. 5

[25] Dingkang Liang, Wei Xu, Yingying Zhu, and Yu Zhou. Focal inverse distance transform maps for crowd localization. *IEEE TMM*, 25:6040–6052, 2023. 3, 5

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 3

[27] Zelin Liu, Xinggang Wang, Cheng Wang, Wenyu Liu, and Xiang Bai. Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth. *arXiv preprint arXiv:2306.05238*, 2023. 5, 6

[28] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. In *ICLR*, 2018. 5

[29] Zhongzhen Luo, Mina Attari, Saeid R. Habibi, and Martin von Mohrenschildt. Online multiple maneuvering vehicle tracking system based on multi-model smooth variable structure filter. *IEEE TITS*, 21(2):603–616, 2020. 2

[30] Zheng Ma and Antoni B. Chan. Counting people crossing a line using integer programming and local features. *IEEE TCSVT*, 26(10):1955–1969, 2016. 2

[31] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. *arXiv preprint arXiv:2302.11813*, 2023. 5, 6, 7

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[33] Zhuojin Pan and Xiujuan Wang. Correlation tracking algorithm based on adaptive template update. In *ICISP*, volume 1, pages 98–101, 2010. 4

[34] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *CVPR*, pages 164–173, 2021. 2

[35] 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. 4

[36] Christopher Rasmussen and Gregory D. Hager. Probabilistic data association methods for tracking complex visual objects. *IEEE TPAMI*, 23(6):560–576, 2001. 2

[37] Christian Ritter, Andrea Imle, Ji Young Lee, Barbara Müller,

Oliver T. Fackler, Ralf Bartenschlager, and Karl Rohr. Two-filter probabilistic data association for tracking of virus particles in fluorescence microscopy images. In *ISBI*, pages 957–960, 2018. 2

[38] Richard Sinkhorn. A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *The Annals of Mathematical Statistics*, 35(2):876 – 879, 1964. 4

[39] Rakshith Sharma Srinivasa, Jaejin Cho, Chouchang Yang, Yashas Malur Saidutta, Ching-Hua Lee, Yilin Shen, and Hongxia Jin. CWCL: Cross-modal transfer with continuously weighted contrastive loss. In *NeurIPS*, 2023. 3

[40] Jason Stauch, Travis Bessell, Mark Rutten, Jason Baldwin, Moriba Jah, and Keric Hill. Joint probabilistic data association and smoothing applied to multiple space object tracking. *JGCD*, 41(1):19–33, 2018. 2

[41] Ramana Sundararaman, Cedric De Almeida Braga, Éric Marchand, and Julien Pettré. Tracking pedestrian heads in dense crowd. In *CVPR*, pages 3865–3875, 2021. 2

[42] Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. Tracking pedestrian heads in dense crowd. In *CVPR*, pages 3865–3875, 2021. 5

[43] Ramana Sundararaman, Cedric De Almeida Braga, Eric Marchand, and Julien Pettre. Tracking pedestrian heads in dense crowd. In *CVPR*, pages 3865–3875, June 2021. 5, 6

[44] Xuan-Thuy Vo, Van-Dung Hoang, Duy-Linh Nguyen, and Kang-Hyun Jo. Pedestrian head detection and tracking via global vision transformer. In *IFCVW*, pages 155–167, 2022. 2, 5, 6

[45] Xingyu Wan, Jinjun Wang, Zhifeng Kong, Qing Zhao, and Shunming Deng. Multi-object tracking using online metric learning with long short-term memory. In *ICIP*, pages 788–792, 2018. 2

[46] Bing Wang, Gang Wang, Kap Luk Chan, and Li Wang. Tracklet association by online target-specific metric learning and coherent dynamics estimation. *IEEE TPAMI*, 39(3):589–602, 2017. 2

[47] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10):3349–3364, 2020. 5

[48] Yu-Hsiang Wang, Jun-Wei Hsieh, Ping-Yang Chen, Ming-Ching Chang, Hung Hin So, and Xin Li. Smiletrack: Similarity learning for occlusion-aware multiple object tracking. *arXiv preprint arXiv:2211.08824*, 2023. 5, 6

[49] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 5

[50] Hefeng Wu, Yafei Hu, Keze Wang, Hanhui Li, Lin Nie, and Hui Cheng. Instance-aware representation learning and association for online multi-person tracking. *PR*, 94:25–34, 2019. 2

[51] 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

[52] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 3

[53] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *ICCV*, pages 5161–5169, 2017. 1, 2

[54] Bo Yang, Chang Huang, and Ram Nevatia. Learning affinities and dependencies for multi-target tracking using a CRF model. In *CVPR*, pages 1233–1240, 2011. 2

[55] Tianyu Yang and Antoni B. Chan. Visual tracking via dynamic memory networks. *IEEE TPAMI*, 43(1):360–374, 2021. 4

[56] 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*, 129:3069–3087, 2021. 2, 5, 6

[57] Zhuoyi Zhao, Hongsheng Li, Rui Zhao, and Xiaogang Wang. Crossing-line crowd counting with two-phase deep neural networks. In *ECCV*, pages 712–726, 2016. 2, 5, 6

[58] 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. 2

[59] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE TPAMI*, 44(11):7380–7399, 2021. 7