Connecting the Complementary-view Videos: Joint Camera Identification and Subject Association

Ruize Han[1], Yiyang Gan[1], Jiacheng Li[1], Feifan Wang[1], Wei Feng[1]†, Song Wang[2]†
[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] Department of Computer Science and Engineering, University of South Carolina, USA
{han_ruize, realgump, threeswords, wff, wfeng}@tju.edu.cn, songwang@cec.sc.edu

Abstract

We attempt to connect the data from complementary views, i.e., top view from drone-mounted cameras in the air, and side view from wearable cameras on the ground. Collaborative analysis of such complementary-view data can facilitate to build the air-ground cooperative visual system for various kinds of applications. This is a very challenging problem due to the large view difference between top and side views. In this paper, we develop a new approach that can simultaneously handle three tasks: i) localizing the side-view camera in the top view; ii) estimating the view direction of the side-view camera; iii) detecting and associating the same subjects on the ground across the complementary views. Our main idea is to explore the spatial position layout of the subjects in two views. In particular, we propose a spatial-aware position representation method to embed the spatial-position distribution of the subjects in different views. We further design a cross-view video collaboration framework composed of a camera identification module and a subject association module to simultaneously perform the above three tasks. We collect a new synthetic dataset consisting of top-view and side-view video sequence pairs for performance evaluation and the experimental results show the effectiveness of the proposed method.

1. Introduction

With the advancement of mobile-camera technologies, human group events such as surprise parties, group games and sports events, are increasingly recorded by various mobile cameras. Wearable cameras, such as GoPro or mobile phone camera, worn by one of the persons (referred to as subjects in this paper) on the ground can provide *side views* of the human group [7, 24, 33]. Unmanned aerial vehicles (UAVs), such as drones in the air, can provide *top views* of the same human group [10]. The video analysis tasks in such two views are both well studied [17, 39, 40]. However, the collaborative analysis of these two views is rarely studied. We can see from Figure 1 that the data collected from these two views *well complement each other* – the top-view video contains no mutual occlusions and well exhibits a global picture and the spatial distribution of the subjects, while the side-view video can capture the detailed appearance, behavior, and activity of subjects of interest in a much closer distance. We believe that their collaborative analysis can help build the *air-ground cooperative visual system* for comprehensive scene understanding, activity analysis, etc.

To achieve this goal, the first challenging problem is to effectively connect these two complementary views. For this we propose to study the following three tasks as shown in Figure 1. **Task I**: Camera location identification – to localize the side-view camera in the top-view video; **Task II**: View direction estimation – to infer the view direction of the side-view camera (in the top view); **Task III**: Cross-view multiple human detection and association – to detect every subject present in each view and identify the same person across the two views.

This is a *very challenging problem* and different from existing works. The biggest challenge lies in that the large (approximately orthogonal) view difference in our setting, which makes *the classical features, e.g., appearance and motion, no longer useful for connecting the two views*. Specifically, Tasks I & II are different from prior works on identifying the first-person camera in a third-person camera [6,35], where the third-person cameras usually adopt the egocentric or surveillance cameras, and their altitudes and angles are similar with the first-person cameras. In this paper, the third-person camera is mounted on a drone, leading to very limited field-of-view (FOV) overlap with the first-person view. This makes prior approaches [6, 35] on modeling the cross-view correspondence fail in our tasks. Task III looks like a specific person re-identification (re-id) problem – for each subject in one view, re-identifying him/her in the other view. However, this is a very challenging person re-id problem because the same subject may show *totally*

† Co-corresponding authors.

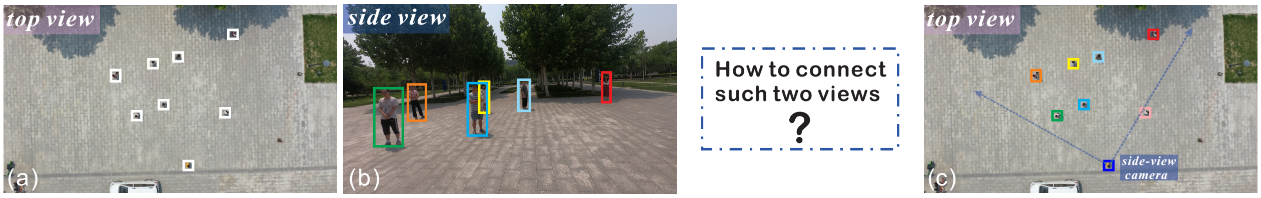

Figure 1. An illustration of the top-view (a) and side-view (b) images. The former is taken by a camera mounted to a drone in the air and the latter is taken by a GoPro worn by a wearer who walks on the ground. To connect such two views, we attempt to answer the following three questions: **Q1**: Who takes the picture (b) in (a)? **Q2**: Where does he/she look at in (a)? **Q3**: Who are the same person across (a) and (b)? True answers of these three questions are shown in (c): the blue box indicates the side-view camera (**Q1**), the two blue arrows indicate view direction of the side-view camera (**Q2**) and identical-color boxes across (b) and (c) indicate the same persons (**Q3**).

different appearance in top and side views, not to mention that the top view of subjects contains *very limited features* by only showing the top of heads and shoulders, as shown in Figure 1.

In this paper, we develop a new approach to explore and leverage the mutual dependence among the above three tasks to solve them simultaneously. Our main idea is to explore the spatial position layout of the subjects in two views. Specifically, we apply a human detection module to detect all the humans in the top and side views, respectively. Based on the detection results, we use a *spatial-aware position representation* to embed the spatial-position distribution of the subjects in different views. To bridge the view gap across the top and side views, we apply the polar transform to the top-view representation for rendering the 360-degree subject distribution appearing in the FOV of the side-view camera. Based on such spatial-aware position representation, we design a *camera identification module* and a *subject association module* to simultaneously infer the side-view camera location and its view direction in the top view, and also match the subjects across the two views. In the experiments, we collect a new large-scale synthetic dataset consisting of rich annotations for model training and performance evaluation. Experimental results verify that the proposed method can effectively handle the proposed three tasks.

The main contributions of this paper are: ❶ This is the first deep model to *jointly handle the above three fundamental tasks* for complementary-view crowded-scene analysis, including the side-view camera localization (Task I), view direction estimation (Task II), and cross-view multi-human detection and association (Task III); ❷ We develop *a new spatial-aware deep framework* including the spatial-aware position representation and complementary-view collaboration network to model and associate the subjects' spatial layout across the complementary views; ❸ We collect *a new large-scale rich-annotation dataset* of top-view and side-view videos for training and evaluating the proposed method. The dataset is released to the public at https://github.com/RuizeHan/DMHA.

2. Related Work

Our Tasks I and II can be regarded as a problem of **identifying the camera holder in third-person cameras**, which has been studied in several works. For example, Fan et al. [6] identify a first-person camera wearer in a third-person video by incorporating spatial and temporal information from the videos of both cameras. Similarly, in [35], subjects are jointly segmented and associated between the synchronized videos captured by the first- and third-person cameras. Differently, in this paper the third-person camera is mounted on a drone and produces top-view images, making cross-view appearance matching very difficult.

As mentioned above, cross-view subject association (Task III) can be treated as a **person re-identification (re-id) problem** [38], which has been widely studied in recent years. Most existing re-id methods can be grouped into two categories: similarity learning and representation learning. The former focuses on learning the similarity metric, e.g., the invariant feature learning-based models [19,27,37], classical metric learning models [16, 20, 23], and deep metric learning models [8, 21, 32]. The latter focuses on feature learning, including low-level visual features such as color, shape, and texture [9, 22], and more recent CNN (Convolutional Neural Network) deep features [4, 25, 31, 41]. These methods assume that all the data are taken from side views, with similar or different view angles, and almost all of these methods are based on appearance matching. In this paper, we attempt to re-identify subjects across top and side views, where appearance matching is not an appropriate choice.

More related to our work is a series of recent works [2, 3, 10, 14, 28] on **collaborative analysis between the top-view and other cameras**. A couple of works [28, 29] propose to determine the location of ground-level images from a large set of top-view aerial images covering the same geographic region, which focus on the large-field localization but not the humans. Some works [10, 13, 14] try to obtain the cross-view human association and tracking by exploring the spatial-aware reasoning. Such works need the predetermined human detection results and an exhaustive search over a very large parameter space. In another series of works [1, 3], by jointly handling a set of egocentric (first-person) side-view videos and a top-view video, a graph-matching-based algorithm is developed to locate all the side-view camera wearers in the top-view video. In [2], the problem is extended to locate not only the camera wearers, but also other side-view subjects in the top-view video. However, this series of methods are based on two assumptions: 1) the top-view camera bears certain slope angle to enable the partial visibility of human body and the use of appearance matching for cross-view association, and 2) the looking-at direction of the side-view camera is the same as the moving direction of the camera wearer. In this paper, we remove these two assumptions that may not be satisfied in real world.

3. Proposed Method

3.1. Overview

We give an overview of the proposed method that mainly contains three stages, as shown in Figure 2. First, we apply a human detection module by applying a CenterNet [5] alike network to get location (heatmaps) of all humans in the top and side views, respectively. Second, we propose to use the human location heatmap to represent the spatial-position distribution of the subjects. To bridge the view gap across the top and side views, we apply the polar transform to the top-view heatmap for rendering the 360-degree subject distribution from the side-view camera (Section 3.2). Based on such spatial-aware subject representation, we design an identification network to simultaneously locate the side-view camera and infer its view direction in the top view (Section 3.3). Finally, we design a cross-view subject association network for matching the subjects across the two views (Section 3.4).

3.2. Spatial-aware Position Representation

Given a pair of images from the top and side views, we first input them into the human detection module, as shown in Figure 2. We use the CNN architecture based on the CenterNet [5] with three heads, i.e., a heatmap head, a box size head and a center offset head. The heatmap head is used for estimating the center positions of the subjects. We can see that the spatial-position layouts of the subjects in the two views are totally different. To bridge this gap, we apply the polar transformation to the top-view heatmap for subject representation.

Top-view subject representation. By examining the complementary-view image pair in Figure 2, we can see that starting from the side-view camera location in the top-view image, the content lying on the same azimuth direction in the top-view image exactly corresponds to a vertical line of the side-view image. This inspires us to apply the polar transformation on the top-view image to build the spatial-aware correspondence between such two views. Specifically, we take the side-view camera location in the top-view image as the origin of polar coordinate and an arbitrary direction, e.g., the south direction, as the 0-degree angle in the polar transform. As shown in Figure 2, the polar transformation between the points (x', y') on the original top-view heatmap $\mathbf{F}^{\mathrm{t}}$ and the target points (x, y) on the expanded heatmap $\tilde{\mathbf{F}}^{\mathrm{t}}$ is defined as

$$\begin{aligned} &\tilde{\mathbf{F}}^{\mathrm{t}}(x,y) = \mathbf{F}^{\mathrm{t}}(x',y'), \\ \text{s.t.}\quad & x' = c_x - r\frac{y}{H}\sin(2\pi\frac{x}{W}), \\ & y' = c_y - r\frac{y}{H}\cos(2\pi\frac{x}{W}), \end{aligned} \tag{1}$$

where (c_x, c_y) locates the origin of polar coordinate on $\mathbf{F}^{\mathrm{t}}$, r is a parameter, W and H denote the width and height of $\tilde{\mathbf{F}}^{\mathrm{t}}$, which are predefined.

Side-view subject representation. Correspondingly, the side-view heatmap can be directly used as the spatial-aware position representation $\mathbf{F}^{\mathrm{s}}$. Given the (appropriately) orthogonal view direction between the top and side views, the vertical lines in the side-view heatmap $\mathbf{F}^{\mathrm{s}}$ correspond to radial lines in the original top-view heatmap $\mathbf{F}^{\mathrm{t}}$, i.e., the vertical lines in the expanded heatmap $\tilde{\mathbf{F}}^{\mathrm{t}}$. Similarly, the transverse lines in $\mathbf{F}^{\mathrm{s}}$ approximately correspond to the concentric circles centered at the polar origin in $\mathbf{F}^{\mathrm{t}}$, i.e., the transverse lines in $\tilde{\mathbf{F}}^{\mathrm{t}}$. With the FOV angle θ of the side-view camera (a fixed camera internal reference, e.g., 90°), the side-view heatmap can be regarded as a part of expanded top-view heatmap. For example, as shown in Figure 2, assuming θ is 90°, the side-view heatmap is a quarter of the top-view heatmap along the width. Here the spatial distributions of the subject position on the side view and the corresponding subregion in top view can be roughly matched.

3.3. Camera Wearer Identification Module

Based on the above subject representations, we propose a two-stage complementary-view collaboration framework to simultaneously address the camera wearer identification and the human detection and association tasks.

View direction searching. First, we discuss the view direction searching of the side-view camera in the top view, that is to match the side-view heatmap $\mathbf{F}^{\mathrm{s}} \in \mathbb{R}^{h\times w}$ with the expanded top-view heatmap $\tilde{\mathbf{F}}^{\mathrm{t}} \in \mathbb{R}^{H\times W}$. As shown in Figure 2, we compress the heatmaps along the y-axis by value accumulation and get $\mathbf{f}^{\mathrm{s}} \in \mathbb{R}^{1\times w}$ and $\tilde{\mathbf{f}}^{\mathrm{t}} \in \mathbb{R}^{1\times W}$. Then we compute the correlation score between them as

$$r_s = \mathbf{f}^{\mathrm{s}} * \tilde{\mathbf{f}}^{\mathrm{t}}_s \in \mathbb{R}, \tag{2}$$

where $*$ denotes the convolution operation, $\tilde{\mathbf{f}}^{\mathrm{t}}_s$ denotes the cropped map from $\tilde{\mathbf{f}}^{\mathrm{t}}$ using a sliding window with the width of w, $s \in \{1, 2, ..., W\}$ denotes the left boundary of the sliding window. Note that, when $s > W - w$, we circularly pad

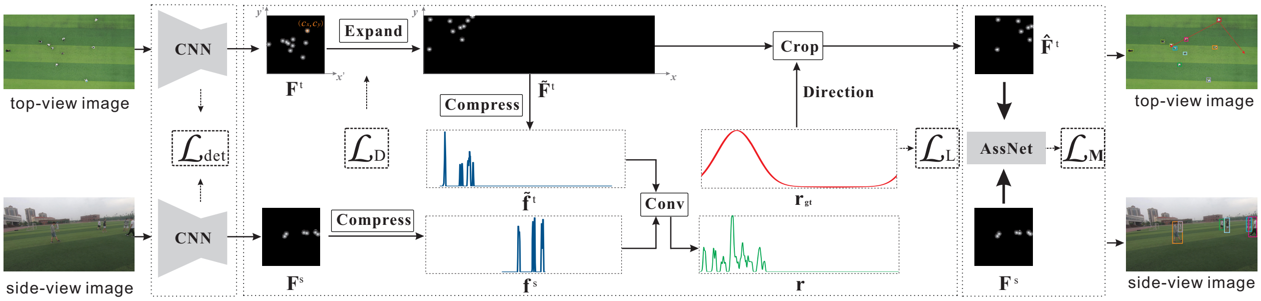

Figure 2. Framework of the proposed method. To zoom in for best view.

the right boundary of $\tilde{\mathbf{f}}^{\text{t}}$ with its left region. Therefore we can get the correlation scores $\mathbf{r} = \{r_s | s \in \{1, 2, ..., W\}\} \in \mathbb{R}^{1\times W}$ corresponding to all view direction candidates.

Camera localization. As discussed above, we take the side-view camera location O in the top-view image as the origin for polar transformation. Actually, we do not know the location O priorly. Therefore, we sample O from the locations of all the subjects $\mathcal{P} = \{P^1, P^2, ..., P^M\}$ in the top view and assume $O \in \mathcal{P}$. In the training process, we take the camera localization as a classification problem. Specifically, if the sampled location is O, it will be taken as a positive sample, while the other sampled locations are taken as the negative ones. In the testing stage, we try all the possible locations from $\mathcal{P}$ and select the predicted camera location with the highest confidence.

Identification network. Based on the above settings, the framework of the camera location and view direction identification network is shown in Figure 2 (middle). We next present the supervisions for training the proposed network. First, for the view direction searching, we use the following direction loss function

$$\mathcal{L}_{\text{D}} = \|\mathbf{r} - \mathbf{r}_{\text{gt}}\|, \tag{3}$$

where $\mathbf{r}$ denotes the view-direction score predicted by Eq. (2) and $\mathbf{r}_{\text{gt}}$ denotes the ground-truth result, i.e., a Gaussian distribution curve as shown in Figure 2. Next, for the camera localization, we use the triplet loss for the camera location prediction

$$\mathcal{L}_{\text{L}} = \log(1 + e^{\tau(\|\mathbf{f}^{\text{s}} - \mathbf{f}^{\text{o}}(\text{o})\| - \|\mathbf{f}^{\text{s}} - \mathbf{f}^{\text{o}}(\text{o}')\|)}), \tag{4}$$

where τ is a pre-set parameter, $\mathbf{f}^{\text{s}}$ is the compressed side-view heatmap in Eq. (2), and $\mathbf{f}^{\text{o}}(\text{o})$ and $\mathbf{f}^{\text{o}}(\text{o}')$ are the compressed heatmaps expanded at true original o and false origin o′, which are taken as the positive/negative samples, respectively.

3.4. Multiple Human Association Module

Subject matching similarity. After the camera localization and view direction searching, we then consider the individual-level subject matching. For that, we first obtain the human bounding box in each view generated by the human detection module in our framework, which are then mapped into the expanded top-view heatmap and side-view heatmap and denoted as P^i for $i \in \{1, 2, ..., M\}$ and Q^j for $j \in \{1, 2, ..., N\}$, respectively. We then measure the similarity between the subjects appearing in top view and side view by their spatial-position layouts.

1) For the x-axis distribution (from left to right in the side-view FOV), we calculate the distance of the x-axis coordinate between each pair of subjects across two views

$$d_x^{i,j} = \text{D}(P_x^i, Q_x^j), \tag{5}$$

where P_x^i, Q_x^i denotes the normalized x-axis coordinate of subject P^i, Q^i, D is a distance measurement function. We then get the similarity matrix $\mathbf{S}_x = \mathbf{1} - [d_x^{i,j}]_{i,j} \in \mathbb{R}^{M\times N}$ between all the subjects across two views.

2) For the y-axis distribution (from near to far in the side-view FOV), we leverage each subject's distance to the camera. Specifically, in the top view, according to the polar transformation discussed above, the y-axis coordinate value directly reflects the distance. In the side view, according to the principle of photography, the distance from the camera can be reflected by the depth of each subject. We calculate the similarity of each pair of subjects along the y-axis

$$d_y^{i,j} = \text{D}(P_y^i, \text{d}(Q^j)), \tag{6}$$

where P_y^i denotes the y-axis coordinate of subject P^i, and $\text{d}(Q_y^j)$ denotes the depth of subject Q^i to the camera. We then get the similarity matrix $\mathbf{S}_y$ similar with that of x-axis. *How to estimate the subject depth in side view?* We try three different ways including ① estimating the depth of each subject by the image depth estimation algorithm; ② using the distance from the bottom of each subject to the bottom of the image; ③ taking the inverse of the human bounding box height. In the experiments, we will compare the results generated in different ways.

Association network. We use bi-directional RNN architecture to construct the association subnetwork inspired by [12, 36]. Given the input similarity matrix $\mathbf{S}_x$ (or $\mathbf{S}_y$), we first reshape it to a vector by following the row-wise order and feed to the first BiRNN, the output of which is then reshaped to a vector by following the column-wise order and then feed it to the second BiRNN. Later, three fully-connected (FC) layers are applied, followed by a sigmoid function to achieve the final matching matrix $\mathbf{M}_x$ (or $\mathbf{M}_y$). We apply the supervised matching loss

$$\mathcal{L}_{\mathrm{M}} = \mathrm{L}_{\mathrm{cro}}(\mathbf{M}_x, \mathbf{M}_{\mathrm{gt}}) + \mathrm{L}_{\mathrm{cro}}(\mathbf{M}_y, \mathbf{M}_{\mathrm{gt}}), \qquad (7)$$

in the subject association network, where $\mathbf{M}_x(\mathbf{M}_y)$ and $\mathbf{M}_{\mathrm{gt}}$ are the predicted matching matrix and the ground truth, respectively. We apply the matrix cross-entropy loss function $\mathrm{L}_{\mathrm{cro}}$ to measure the consistency between two matrices.

4. Implementation Details

Network training. The total loss function of the whole framework is defined as the summation of the detection loss $\mathcal{L}_{\mathrm{Det}}$, direction loss $\mathcal{L}_{\mathrm{D}}$, localization loss $\mathcal{L}_{\mathrm{L}}$ and matching loss $\mathcal{L}_{\mathrm{M}}$ as defined above, i.e., $\mathcal{L} = \mathcal{L}_{\mathrm{Det}} + \mathcal{L}_{\mathrm{D}} + \mathcal{L}_{\mathrm{L}} + \mathcal{L}_{\mathrm{M}}$. In some cases, the annotations of the location and (especially) the view direction of the camera are hard to require. For this situation, the proposed method can also implement without $\mathcal{L}_{\mathrm{D}}$ or $\mathcal{L}_{\mathrm{L}}$ or both of them, which shows the generality of the proposed method and the corresponding results are discussed in Section 5.3. In the detection module, we apply the network architecture used in CenterNet [5] as the backbone. In the experiments, we resize both the width and height of $\mathbf{F}^{\mathrm{t}}$, denoted as w and h, as 128. In Eq. (1), we take $r = \frac{w}{2}$. We set $H = h, W = \lambda w$ as the the width and height of $\tilde{\mathbf{F}}^{\mathrm{t}}$, and set $\lambda = 4$, since the FOV angle θ of the side-view camera is 90°. We set τ in Eq. (4) as 10^2. We use Pytorch backend for implementing the proposed network and run on a computer with RTX 3090 GPU.

Network inference. We then elaborate on the inference stage of the proposed method. First, we use the convolution to achieve view direction searching. Specifically, we apply the convolution operation on $\mathbf{f}^{\mathrm{s}}$ and $\tilde{\mathbf{f}}^{\mathrm{t}}$ as Eq. (2) to get $\mathbf{r} \in \mathbb{R}^W$ as the response score. We take the peak value on the response score to get the view direction. For camera localization, we try all the possible locations $\mathcal{P} = \{P^1, P^2, ..., P^M\}$ as the origin of polar transformation and calculate the corresponding localization errors $\|\mathbf{f}^{\mathrm{s}} - \mathbf{f}^{\mathrm{o}}(P)\|$ as defined in Eq. (4) to select the predicted camera location with the minimum error. For the subject matching task, we merge the predicted $\mathbf{M}_x$ and $\mathbf{M}_y$ by averaging them and get $\mathbf{M}$, we then apply the Hungarian algorithm [15] on the predicted soft matching matrix $\mathbf{M}$ to transform the output into a hard (binary) assignment matrix $\mathbf{A}$ as the final subject association result.

5. Experiments

5.1. Dataset

Synthetic dataset. We do not find available datasets containing complementary top- and side-view videos with the full annotations of side-view camera location, view direction and cross-view subject association. Especially for the side-view camera view direction, no matter using the auxiliary hardware instruments or manual post-annotation, it is very hard to be accurately obtained in real-world data collection. Thus, we consider building a synthetic dataset.

• *Controllable data collection.* We leverage a 3D modeling engine Unity [26] to render the background. We further apply an open-source toolkit PersonX [30] to model the humans appearing in the synthetic videos. We generate the complementary-view video pair by using a top-view camera from a high altitude, and a side-view camera that is mounted on the head of a subject in the scene. Benefiting from the virtual environment, we can control a series of setups.

• *Diverse settings.* We apply five common outdoor surveillance scenes like the city street, campus, and stadium, where we select 10 different sites for video collection. We also include day and night scenes with various illuminations. The number of subjects in each video is set in the range of 5 – 25, which are randomly selected from 1,000 3D human models. All the subjects are controlled to walk/stand freely in the scene without specific requirements. The altitude of the top-view camera is set as 15 – 20 meters, which looks nearly vertically down to the ground that can cover all/most subjects in the scene. The side-view camera is still or moving with the movement of the camera wearer, which includes random walking, and head rotating/pitching. We do not require all the subjects to be visible in the side-view camera, but we make the FOV of side-view camera to cover most of the subjects. This is also common in the surveillance scenarios. The field-of-view angle of the side-view camera is set as 90° following many real-world mobile cameras.

• *Large scale.* We generate 108 videos (54 video pairs) with the length varying from 500 to 1,500 frames, which, in total, includes 84,800 frames with over one million subject bounding boxes. We split the dataset into training and testing sets by 2 : 1, i.e., 36 and 18 videos, respectively.

• *Rich and accurate annotations.* All the necessary annotations used in this problem including the side-view camera location, view direction (in top-view video), and human bounding boxes with temporal and cross-view ID numbers, can be accurately obtained in our setting.

Real-world dataset. We also include a real-world dataset [11] in the experiments. Specifically, this dataset is collected by GoPro camera (mounted over wearer's head) to take side-view videos and a UAV to take top-view videos. The dataset includes 15 video pairs with the length varying from 600 to 1,200 frames, which are taken at five different sites with various backgrounds. The number of subjects in each video varies from 3 to 14. We split the dataset into training and testing datasets, with 8 and 7 video pairs, respectively. The subjects are manually annotated in the forms of bounding boxes with ID numbers: the same subject across the two views is annotated with the same ID number. Note that, this manual labeling is very labor-intensive given the difficulty to identify subjects in the top-view videos. The dataset only provides the annotations for camera wear localization and cross-view subject association but not the view direction, which, actually, is almost impossible to be accurately annotated in the real-world dataset.

5.2. Setup

Evaluation metrics. To comprehensively evaluate the proposed method, we define the following metrics.

Metric-I: We first evaluate the accuracy of *side-view camera localization*. For each frame, given the predicted and ground-truth camera wearer O^{p} and O^{g} (in terms of human bounding box), respectively, we take the localization result to be true if the Intersection over Union (IoU) of O^{p} and O^{g} is larger than $\frac{1}{2}$. We then rank all the detected subjects in the top view based on its prediction score to be the camera wearer generated by the algorithms and evaluate the top-κ accuracy, which denotes the true camera location is among the top κ-proportion of the ranked detected subjects.

Metric-II: We also evaluate the *side-view-camera view direction* estimation. Given the predicted and ground-truth view direction V^{p} and V^{g} (in terms of angle within $[0, 2\pi)$), respectively, we first calculate the view direction error as $\gamma = |V^{\mathrm{p}} - V^{\mathrm{g}}|$. We define the accuracy δ^{α} as the percentage of predicted view directions satisfying that $\gamma \leq \alpha$.

Metric-III: We finally evaluate *cross-view multiple human association* results. Specifically, we use the precision and recall scores for the cross-view subject association evaluation, which are calculated by the number of correctly matched subjects over all the predicted or ground-truth ones, respectively. We also compute the F_1 score as the metric. We further use the multi-human association accuracy $\mathrm{MHAA} = 1 - \left(\frac{\sum_t \mathrm{fn}_t + \mathrm{fp}_t + 2\mathrm{mme}_t}{\sum_t \mathrm{g}_t}\right)$, where fn_t, fp_t, and mme_t are the numbers of false negatives, false positives, and mismatch pairs of cross-view subject matchings at time t, respectively, and g_t is the total number of subjects in both the top and side views at time t [10]. Note that, Metric-III is a comprehensive metric evaluating the performance of both the human detection and association. We do not separately evaluate the single-view human detection precision because it is not the main purpose of this work.

Comparison methods. We did not find available methods with code that can directly handle our problem, especially for the proposed Task I and Task II. Specifically, previous works [6, 35] all use the congeneric first-person and third-person cameras with common height and FOV. Differently, in this paper the top view makes the cross-view appearance and motion, the most important features in previous works [6,35], very difficult to match for camera identification. Moreover, given the unreachable annotations for the view direction, there is no previous works to estimate and evaluate the view direction of a first-person-view camera from a third-person view. Task III is also different from most existing works that focus on matching the subjects with similar appearance/motion features. Even so, we still try to include more related approaches with some modifications for the comparison of subject association.

- *MOT* : We first use a top-rank appearance-motion-based multiple object tracking (MOT) algorithm TraDes [34] for comparison. Specifically, we *manually associate* the subjects between top and side views only on the frames when each subject first appears in the video. We then track all the subjects in each video by TraDes, respectively, and finally using the tracking results to propagate the subject association to later frames.
- *Re-id*: The cross-view subject association task is similar to the appearance-matching-based person re-id methods. So, we choose a state-of-the-art person re-id approach [4, 25] for the cross-view subject association. We apply the re-id network to extract the feature of each subject and calculate the similarities among the subjects in two views, and then choose the matched subject pairs between different views with the maximum similarity.
- *MHA* : The most similar work to our task is the one in [10, 13] for cross-view multi-human association, which constructs a cost function to measure the similarity across two views with large view difference.

Note that, the above three methods need the human detection as input. For a fair comparison, in the experiments, they all use the detection results generated by our method.

- *Hungarian* + **S**: We directly apply the Hungarian algorithm [15] on the similarity matrix $\mathbf{S}$ (average of $\mathbf{S}_x$ and $\mathbf{S}_y$) to get the assignment matrix $\mathbf{A}$ without using the proposed association network to predict $\mathbf{M}$.

5.3. Camera Identification Results

We first evaluate the performance of the camera wearer localization (Task I) and the view direction estimation (Task II). To provide more comprehensive comparisons for them, we apply two baseline approaches used for the task III, i.e., *Re-id* and *MHA*, to handle Tasks I and II. To be specific, with the human detection and the subject association results (Task III), we localize the side-view camera in the top view by searching each subject to identify its localization and view direction that covers most the (associated) subjects except the camera wearer.

Ablation study. We consider the following variations of our method to verify some key components.

- w/o $\mathcal{L}_{\mathrm{D}}$ / $\mathcal{L}_{\mathrm{L}}$: Remove the direction or localization loss in camera wearer identification. Note that, the real-world dataset does not have the direction annotations thus we do not use $\mathcal{L}_{\mathrm{D}}$ in our method.
- w/o compress: Do not compress the heatmap into a vector before applying the correlation operation in Eq. (2).

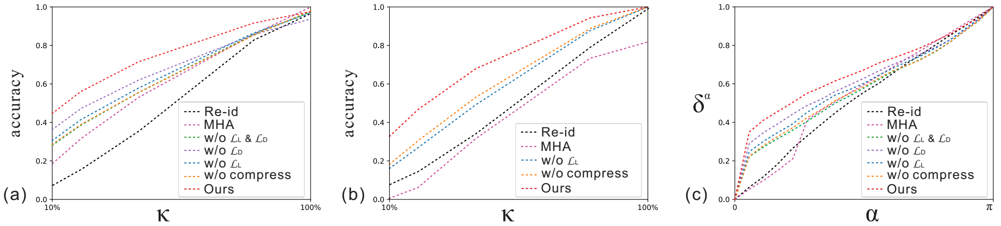

Figure 3. Comparative results of different variations of our method for Tasks I (a,b) and Task II (c).

Table 1. AUC scores of different variations of our method. (%)

Method	Location (Syn.)	Location (Real)	Direction
Re-id [4, 25]	61.86	60.61	56.00
MHA [10]	69.95	52.58	57.90
w/o $\mathcal{L}_\mathrm{D}$	73.72	-	64.17
w/o $\mathcal{L}_\mathrm{L}$	72.69	69.21	62.09
w/o $\mathcal{L}_\mathrm{D}$ & $\mathcal{L}_\mathrm{L}$	71.49	-	59.65
w/o compress	71.51	71.02	60.29
Ours	**80.50**	**79.26**	**68.34**

For Task I, following the Metric-I as discussed above, we rank all the detected subjects in the top view based on their predicted possibility to be the camera wearer. We then draw the accuracy CMC (Cumulative Matching Characteristics) curves to evaluate the camera-wearer's detection accuracy in the top view. Figure 3a and Figure 3b show the accuracy CMC curves generated by the different variations of our method on the synthetic dataset and real-world dataset, respectively. We can see that our method outperforms the comparative methods in Tasks I and II. For Task I, we can also see that the proposed losses, including $\mathcal{L}_\mathrm{D}$ and $\mathcal{L}_\mathrm{L}$, and the compress strategy are useful for the camera localization task. For quantitative evaluation, we calculate the AUC (Area Under The Curve) score of the CMC curve as shown in the first two columns of Table 1. For Task II, following the Metric-II, we draw the accuracy curve of δ^α along different settings of the threshold α on the synthetic dataset as shown in Figure 3c. The last column in Table 1 shows the AUC scores of the δ^α curves. We can see similar results as for Task I that the proposed components are effective. For in-depth analysis, the proposed method can generate the acceptable camera localization and view direction prediction results without the corresponding supervisions, which demonstrates the robustness of the proposed framework that it can achieve these two tasks in an unsupervised manner.

5.4. Subject Association Results

Comparative results. We then evaluate the subject association results (Task III). In order to better evaluate the association task, we compute the performance of the proposed method given the ground-truth camera location (but not the view direction) as in [2, 10], which is relatively easy to be obtained in the real-world application. As shown in Table 2, we can see that although we give matching labels at the initial frame, the state-of-the-art MOT method TraDes still produces a poor performance in our task. The reason might be that, tracking error of a subject in one frame may cause the association error of this subject in all the frame after. Similarly, the performance generated by the human re-id method is also not good. This is because the existing re-id method relies heavily on the appearance feature, which, however, is not consistent across the top and side views. The method MHA provides an acceptable result, particularly in its self-proposed real-world dataset. Compared with them, the proposed method produces better results on both the synthetic and real-world datasets. Besides, we can see that the comparative method using the similarity matrix **S** and the Hungarian [15] algorithm also performs worse than ours with the assignment network. This verifies the effectiveness of the proposed assignment network, which can handle the similarity measurement errors in **S**.

Ablation study. We also consider several variations of the proposed method.

- w/o x (y) : We remove the matching similarity provided by the x coordinate (y coordinate), respectively.
- w depth / bottom: We use the ① estimated depth of each subject by [18] or ② the distance from the bottom of each subject to the bottom of the image, for the y-axis distribution in Eq. (6).

As shown at the bottom of Table 2, we can see that the subject matching similarity with only the x-axis distribution can provide an acceptable performance. In contrast, the method with only the y-axis distribution performs not very well. This is because the subjects' distributions along x-axis in two views are aligned, given the predicted view direction of the side-view camera. But the scale of the y-axis distributions reflected in the top and side views are non-uniform. Anyway, the final version of our method integrating both of them provides better performance than using any one of them. This demonstrates that the x-axis and y-axis distributions can complement each other in our method. We can also see that the human depth generated by a depth estimation method [18] and calculated by the bottom distance performs not well as the usage of human bounding box height in our method. The reason might be the human depth estimation using [18] is not accurate enough for this problem and the using of bottom distance is easily to be influenced by the rolling of the side-view camera.

Table 2. Comparative results of different methods and different variations of our method. (%)

Method	Synthetic dataset				Real-world dataset			
	Precision	Recall	F_1 score	MHAA	Precision	Recall	F_1 score	MHAA
TraDes [34]	15.34	3.79	6.07	19.03	18.68	6.72	9.89	28.51
Re-id [4, 25]	35.28	20.50	25.93	30.89	26.37	15.33	19.39	33.21
MHA [10]	49.99	41.06	45.09	45.42	73.14	69.04	71.03	67.70
Hungarian [15]	57.44	59.64	58.28	52.07	67.17	75.77	70.76	73.21
w x (w/o y)	58.33	60.71	59.24	57.11	69.54	80.36	74.04	79.43
w y (w/o x)	39.94	41.97	40.73	34.13	40.67	48.33	43.80	47.97
w depth [18]	59.63	61.81	60.46	57.40	70.41	81.11	74.90	79.76
w bottom	63.23	66.05	64.32	60.58	70.47	80.85	74.82	79.16
Ours	**67.06**	**69.91**	**68.16**	**66.07**	**72.05**	**83.50**	**76.81**	**80.80**

Cross-domain testing. Clearly, the view direction annotation for the real-world data is quite hard. Even using the gyroscopes integrated in the smart phones and cameras, it cannot solve this problem either, e.g., external disturbances produces random drift error all the time. This way, we test and evaluate the results on the real-world data using the model trained on the synthetic data, to evaluate the generalization ability of our method. As discussed above, the view direction on real-world videos can not be acquired. Therefore, we evaluate cross-view subject association performance as shown in Table 3. We can see that, although with some accuracy drop, the cross-domain testing still provides acceptable performance. Note that, we directly apply the saved model trained on the synthetic training dataset for real-world data testing without any extra modification. We believe the performance can be better by integrating some techniques for the cross-domain adaption or synthetic data generation. From this point, this paper provides a new insight that using synthetic data may help detect the (unmeasurable) first-person view direction in real-world scene.

Table 3. Cross-domain evaluation of our method. (%)

Method	Real-world dataset			
	Precision	Recall	F_1 score	MHAA
Ours (Cross-domain)	52.76	65.88	57.85	70.62

6. Discussion

Limitation. 1) We assume the side-view camera always locates in the FOV of the top-view camera, which may not be always satisfied in practice. 2) We do not use the temporal information in the video. While it has the advantage to be applicable to single image pairs, videos can provide more information, e.g., the temporal consistency, for performance improvement. We also show some special cases to discuss the limitation in the supplementary material.

Application. Actually, this work handles a new problem setting of air-ground cooperative camera system. Note that, in an outdoor scenario without pre-installed cameras, it may be unpractical to quickly setup the traditional fixed cameras for surveillance. This way, the proposed camera system can be applied: cameras on a drone (top view) and worn by several law enforcement officials on the ground (side views) can be deployed, with the proposed association, for collaborative localization, tracking, and human activity recognition, etc. The complementary-view camera configuration can provide outdoor surveillance with much better coverage and flexibility since the top and side views well complement each other, where the top view provides a global picture of the whole scene but lacks details, while the side views provide local details of subjects with frequent occlusions. With the advancement of mobile-camera technologies, the benefit of the collaborative analysis of such cameras will also increase, with many potential applications in video surveillance, e.g., human group activity recognition [39], important person detection [14], and sport scene understanding, e.g., player positioning analysis in football games.

7. Conclusion

In this paper, we have studied a new problem for complementary-view video collaborative analysis. For that, we developed a new approach that can simultaneously handle three tasks – camera wearer location, view direction estimation and cross-complementary-view multiple human detection and association. Specifically, we have proposed a spatial-aware position representation method to embed the spatial distribution of the subjects and designed a camera identification and subject matching network to simultaneously perform the above three tasks. We also built a new synthetic dataset with rich annotations for the proposed problem. Experimental results on both the synthetic dataset and a real-world dataset are very promising. In the future, we plan to integrate the temporal information of the videos into our framework for further improving the performance.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grants U1803264, 62072334, and the Tianjin Research Innovation Project for Postgraduate Students under Grant 2021YJSB174, and the Natural Science Foundation of Tianjin under Grant 18JCYBJC15200.

References

[1] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *ECCV*, 2016. 3

[2] Shervin Ardeshir and Ali Borji. Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *ECCV*, 2018. 2, 3, 7

[3] Shervin Ardeshir and Ali Borji. Egocentric meets top-view. *IEEE TPAMI*, 41(6):1353–1366, 2019. 2, 3

[4] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *ICCV*, 2019. 2, 6, 7, 8

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, 2019. 3, 5

[6] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *CVPR*, 2017. 1, 2, 6

[7] Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *ACM MM*, 2021. 1

[8] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020. 2

[9] Douglas Gray and Tao Hai. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 2

[10] Ruize Han, Wei Feng, Yujun Zhang, Jiewen Zhao, and Song Wang. Multiple human association and tracking from egocentric and complementary top views. *IEEE TPAMI*, 2021. 1, 2, 6, 7, 8

[11] Ruize Han, Wei Feng, Jiewen Zhao, Zicheng Niu, Yujun Zhang, Liang Wan, and Song Wang. Complementary-view multiple human tracking. In *AAAI*, 2020. 5

[12] Ruize Han, Yun Wang, Haomin Yan, Wei Feng, and Song Wang. Multi-view multi-human association with deep assignment network. *IEEE TIP*, 31:1830–1840, 2022. 4

[13] Ruize Han, Yujun Zhang, Wei Feng, Chenxing Gong, Xiaoyu Zhang, Jiewen Zhao, Liang Wan, and Song Wang. Multiple human association between top and horizontal views by matching subjects' spatial distributions. In *arXiv*, 2019. 2, 6

[14] Ruize Han, Jiewen Zhao, Wei Feng, Yiyang Gan, Liang Wan, and Song Wang. Complementary-view co-interest person detection. In *ACM MM*, 2020. 2, 8

[15] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 5, 6, 7, 8

[16] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 2

[17] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *CVPR*, 2021. 1

[18] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 7, 8

[19] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 2

[20] Giuseppe Lisanti, Iacopo Masi, Andrew D. Bagdanov, and Alberto Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE TPAMI*, 37(8):1629–1642, 2015. 2

[21] K. B. Low and U. U. Sheikh. Learning hierarchical representation using siamese convolution neural network for human re-identification. In *ICDIM*, 2015. 2

[22] Bingpeng Ma, Su Yu, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV*, 2012. 2

[23] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, 2015. 2

[24] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. 3d gaze concurrences from head-mounted cameras. In *NeurIPS*, 2012. 1

[25] Rodolfo Quispe and Helio Pedrini. Top-db-net: Top dropblock for activation enhancement in person re-identification. In *ICPR*, 2020. 2, 6, 7, 8

[26] John Riccitiello. John riccitiello sets out to identify the engine of growth for unity technologies (interview). *VentureBeat. Interview with Dean Takahashi. Retrieved January*, 18(3), 2015. 5

[27] Zhao Rui, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 2

[28] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *NeurIPS*, 2019. 2

[29] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *CVPR*, 2020. 2

[30] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *CVPR*, 2020. 5

[31] Xiao Tong, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 2

[32] Rahul Rama Varior, Shuai Bing, Jiwen Lu, Xu Dong, and Wang Gang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016. 2

[33] Minh Vo, Ersin Yumer, Kalyan Sunkavalli, Sunil Hadap, Yaser Sheikh, and Srinivasa G. Narasimhan. Self-supervised multi-view person association and its applications. *IEEE TPAMI*, 43(8):2794–2808, 2021. 1

[34] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *CVPR*, 2021. 6, 8

[35] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first- and third-person videos. In *ECCV*, 2018. 1, 2, 6

[36] Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixé, and Xavier Alameda-Pineda. How to train your deep multi-object tracker. In *CVPR*, 2020. 4

[37] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Yi Dong, and Stan Z. Li. Salient color names for person re-identification. In *ECCV*, 2014. 2

[38] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C.H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, 2021. 2

[39] Jiewen Zhao, Ruize Han, Yiyang Gan, Liang Wan, Wei Feng, and Song Wang. Human identification and interaction detection in cross-view multi-person videos with wearable cameras. In *ACM MM*, 2020. 1, 8

[40] Kang Zheng, Xiaochuan Fan, Yuewei Lin, Hao Guo, Hongkai Yu, Dazhou Guo, and Song Wang. Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras. In *ICCV*, 2017. 1

[41] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *CVPR*, 2017. 2