International Journal of Computer Vision (2023) 131:1106–1121
https://doi.org/10.1007/s11263-022-01744-z

Relating View Directions of Complementary-View Mobile Cameras via the Human Shadow

Ruize Han[1] · Yiyang Gan[1,2] · Likai Wang[1] · Nan Li[1] · Wei Feng[1] · Song Wang[3]

Received: 14 February 2022 / Accepted: 14 December 2022 / Published online: 11 January 2023

Abstract

The potential of video surveillance can be further explored by using mobile cameras. Drone-mounted cameras at a high altitude can provide top views of a scene from a global perspective while cameras worn by people on the ground can provide first-person views of the same scene with more local details. To relate these two views for collaborative analysis, we propose to localize the field of view of the first-person-view cameras in the global top view. This is a very challenging problem due to their large view differences and indeterminate camera motions. In this work, we explore the use of sunlight direction as a bridge to relate the two views. Specifically, we design a shadow-direction-aware network to simultaneously locate the shadow vanishing point in the first-person view as well as the shadow direction in the top view. Then we apply multi-view geometry to estimate the yaw and pitch angles of the first-person-view camera in the top view. We build a new synthetic dataset consisting of top-view and first-person-view image pairs for performance evaluation. Quantitative results on this synthetic dataset show the superiority of our method compared with the existing methods, which achieve the view angle estimation errors of 1.61° (pitch angle) and 15.13° (yaw angle), respectively. The qualitative results on real images also show the effectiveness of the proposed method.

Keywords Complementary view · Mobile camera · Camera calibration · Shadow vanishing point

Communicated by Stefano Mattoccia.

Ruize Han and Yiyang Gan contributed equally to this work

✉ Wei Feng
wfeng@tju.edu.cn

✉ Song Wang
songwang@cec.sc.edu

Ruize Han
han_ruize@tju.edu.cn

Yiyang Gan
realgump@tju.edu.cn

Likai Wang
kkww@tju.edu.cn

Nan Li
linan94@tju.edu.cn

[1] School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China

[2] Meituan, Beijing 100083, China

[3] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

1 Introduction

Various mobile cameras can capture richer visual information for high-performance surveillance (Han et al., 2022; Liu et al., 2016). On one hand, cameras worn by people on the ground, such as phone cameras and GoPro, can conveniently approach and record the nearby people and events from a *first-person view* (Lin et al., 2015; Zhao et al., 2020; Zheng et al., 2017). On the other hand, aerial cameras, such as those mounted to a drone, can capture a bird's-eye *top view* of the scene from high altitude, which is also widely used in many civil and military scenarios (Barekatain et al., 2017; Li et al., 2021; Perera et al., 2019; Singh et al., 2018; Zhang et al., 2020). As presented in Han et al. (2022), the *complementary view* is defined as the combination of a top view from the camera in air, and a first-person view from a camera on the ground. Note that, the top view with a high altitude can provide the global picture of a crowd scene, while the first-person view provides the local details of interest with a flexible field of view (FOV). Recent research has shown that the collaborative analysis of these two complementary views can significantly enhance the capability of video surveil-

lance (Ardeshir & Borji, 2018a, b; Han et al., 2020b). For example, in an outdoor scenario without pre-installed cameras, we can associate the videos taken by the cameras on a drone (top view) and worn by several law enforcement officials (first-person views) for collaborative tracking (Han et al., 2020a), individual/group activity recognition (Zhao et al., 2020), important person identification (Han et al., 2020b) and anomaly detection, etc. The above setting also *becomes more and more available in practice due to the widespread use of drones and various wearable cameras*.

To establish such *complementary-view mobile camera system* and further explore comprehensive information, we need to first build certain cross-view correspondence at the pixel, region, structure, or object levels. Previous works try to address or simplify this problem by making certain assumptions, e.g., consistent view and motion directions of the camera wearer (Ardeshir & Borji, 2018a, b), or zero pitch angle of the first-person view camera (Han et al., 2020b), but many of these assumptions may not hold in practice. While a full relative pose calibration between the two complementary-view cameras can thoroughly solve this problem, it is very difficult given the indeterminate camera motions and the significant view difference. In this work, we attempt to address a *weaker version of calibration – localizing the field-of-view (FOV) of the first-person view camera in the global top view* by estimating its relative yaw and pitch angles. Here we do not consider the roll angle of the first-person view camera, since the head tilt is not common for a camera wearer in real-world applications. This weak camera calibration can relate the two complementary-view directions and help address many important surveillance tasks, such as cross-view person identification (Han et al., 2019, 2020a) and co-attention person detection (Han et al., 2020b), as discussed in later sections.

Relating the two complementary-view directions is still a very challenging problem—the top view direction is largely perpendicular to the ground while the first-person view can be parallel to the ground, as shown in Fig. 1a–b. Existing methods are mostly based on key-point detection and matching, followed by estimating a multi-view geometry transform. However, key-point features vary significantly and usually cannot be correctly matched across the first-person and top views. Identifying human body joints as key points is also infeasible—in top view, each person on the ground can be very small and only his/her head top and two shoulders are visible. Gyroscopes integrated in the smart phones and cameras cannot be used to solve the proposed problem either, e.g., external disturbances produce random drift error in yaw angle measurement all the time, especially when a magnetic source is nearby.

In this work, we propose a new approach of leveraging the sunlight and the human shadow vanishing point to address the above problem. The *vanishing point* is generally defined as a point in the perspective drawing onto which parallel lines appear to converge. In this work, we specifically adopt the shadow vanishing point, i.e., the intersection point of lines along the human shadows in the image taken by the camera. The top view is nearly vertical to the ground, on which the shadows keep parallel and the (shadow) vanishing point is treated as intersecting at a point at infinity. In the first-person view, the shadow vanishing point may be located within or outside the image perimeter, and actually even at infinity in the degenerate cases. In specific, as shown in Fig. 1d, given sunlight source at infinity and largely vertical top-view direction, shadows of different people on the ground are parallel to each other and have the same direction in the top view. In the first-person view, the lines passing through these shadows intersect at a common *(shadow) vanishing point*, as shown in Fig. 1c. Based on this, we propose to use the sunlight direction as a bridge to relate the first-person and the top views. More specifically, after detecting shadows in both views, we propose a simple yet effective Shadow Vanishing Point Detection Network (SVPN) to simultaneously get the shadow vanishing point in the first-person-view image and the shadow direction in the top-view image. With them, we build a geometric transformation based on the multi-view geometry (Hartley & Zisserman, 2003) to estimate the first-person view direction relative to the top view.

In the proposed method, we use a deep neural network for the shadow vanishing point/direction detection task. This is because the deep learning based method has shown superior performance in various detection tasks. After that, we use a classical geometric model for the (weak) camera calibration task, since in which the deep-learned solutions are currently not preferable over classical ones, especially for the complementary view in our problem with very limited overlapped FOV for deep feature extraction. Note that, the proposed method is only applicable to the scenes with the presence of multiple ($\geq$2) people and sunlight that casts shadows, as

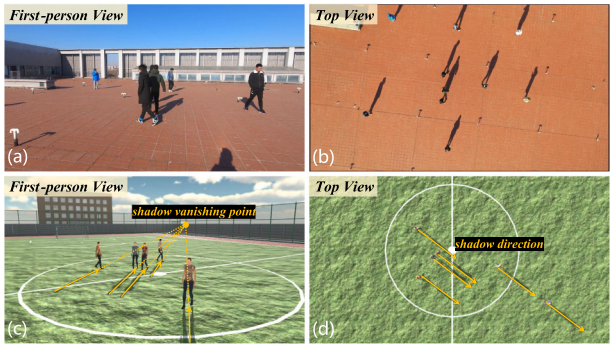


Fig. 1 An illustration of the proposed problem of relating **a** the first-person view and **b** the synchronous top view of a real-world scene. Illustration of the shadow vanishing point localization in the first-person view (**c**) and shadow direction estimation in the top view (**d**) in the simulation environment

shown in Fig. 1. Other light interference shall be weak by not producing shadows, and other objects and their corresponding shadows are allowed in the scene. With these restrictions, *the applicable scenes are actually quite common for daytime outdoor video surveillance* – the goal of many surveillance tasks is to detect, track and recognize the activities in multi-person scenes.

The main contributions of this work are:

① We study a new problem of relating the view directions of two complementary-view mobile cameras: one for the top view from a high altitude and the other for the first-person view on the ground. Although previous works have tried to estimate the camera pose with large view difference or little FOV (field of view) overlap. There are few existing work studies to relate the camera pose with approximately orthogonal view angles and such far distance (a few tens of meters) in this work. To the best of our knowledge, this work is the first to specifically study this new and challenging problem, which is a fundamental problem for the complementary-view video collaborative analysis.

② We provide a new insight for camera calibration with large view difference. Specifically, the proposed method explores the sunlight direction (a common and stable natural phenomenon) as the cue to relate the complementary views. We novelly adopt the (human shadow) vanishing point detected in both views to build the multi-view geometric transformation. We also establish a framework based on the above insight to solve the proposed problem, in which we integrate a deep network based vanishing point detection module and a multi-view geometry based camera relating module. This framework integrates both the generalization of the deep network for detection task and the theory guarantee of the classical geometry for camera pose estimation task.

③ We have built a benchmark including the controllable shadow generation tool ShadowX, dataset, annotation, and evaluation metrics for this problem, which are released at https://github.com/realgump/CVCR. We hope that these resources can establish a research foundation for the community to study the proposed new yet important problem.

2 Related Work

Cameras Extrinsic Calibration Applications with multiple large-view-difference cameras usually require extrinsic calibration to determine their accurate relative poses and many methods have been developed to solve this problem (Miraldo et al., 2015; Guan et al., 2021). In Liu et al. (2014), high-precision measuring devices, such as laser trackers and laser range finders, are adopted to help the extrinsic calibration of cameras. In Dong et al. (2016), Birdal et al. (2016), visual measuring instruments are employed to bridge the gap between the FOVs of different cameras. These methods require additional devices which may not be available in many applications. In Micusik (2011), Censi et al. (2013), different structure from motion (SfM) (Schonberger & Frahm, 2016) methods are developed to track the movement of targets and establish the FOV relationship between different cameras. These methods have difficulty to handle the complementary views discussed in this work, especially the top view where subjects are of very small size with little appearance details.

Vanishing Point Detection Vanishing point detection is a fundamental problem in computer vision (Magee & Aggarwal, 1984; Yang et al., 2016; Zhai et al., 2016). After the initial work proposed by Barnard (1983), various methods have been developed for finding different vanishing points in 2D images (Antunes & Barreto, 2013; Barinova et al., 2010; Coughlan & Yuille, 1999; Kluger et al., 2017; Lezama et al., 2014; Schindler & Dellaert, 2004; Vedaldi & Zisserman, 2012; Wildenauer & Hanbury, 2012a). One popular way for vanishing point detection is to cluster the line segments followed by different refinement procedures (Lezama et al., 2014; Schindler & Dellaert, 2004; Tardif, 2009; Wildenauer & Hanbury, 2012b). Many clustering algorithms include RANSAC (Bolles & Fischler, 1981), J-linkage (Tardif, 2009), Hough transform (Hough, 1959), and EM (Kogecka & Zhang, 2002) have been explored for solving this problem. Recently, deep-learning methods have shown great success in vanishing point detection (Borji, 2016). The key idea is to extract the global image context using a deep convolutional network and then use it to help select vanishing points from a set of candidates under consideration. Kluger et al. (2017) presented a CNN-based approach for detecting vanishing points from a Gaussian sphere representation. Lee et al. (2017) proposed a unified end-to-end network that jointly handles the lane and road marking detection. Recently, Zhou et al. (2019) present a simple yet effective deep network with geometry-inspired convolutional operators for detecting vanishing points in images. Several works also leverage the sun light or vanishing points for vision based applications. Balcı and Güdükbay (2017) estimate and utilize the sun position based on shadow length and utilized this estimation to insert synthetic objects into a real video with their shadows. Doğan et al. (2021) utilize the vanishing points and an image-based camera configuration method to automatically reconstruct navigable regions in a crowd video to augment virtual agents seamlessly into the real video. While all these methods focus on vanishing points of parallel line structures like lanes, roads, and buildings, in this work, we detect and use the vanishing point of person shadows, which may show more complex shape.

Complementary-View Camera Collaboration Recently, collaborative analysis of multiple videos taken from the top view and the first-person view has drawn much attention in the vision community (Ardeshir & Borji, 2016, 2018a, b; Ardeshir et al., 2016; Han et al., 2019, 2020a, b). Ardeshir and Borji (2016), Ardeshir and Borji (2018a) propose to identify the egocentric camera wearer in the top view using synchronized video pairs. In Ardeshir and Borji (2018b), it is extended to simultaneously identify both the egocentric-camera wearer and the other subjects in top views. These works require the view direction of the first-person-view camera to be consistent with moving direction of the camera wearer, which may not hold when the wearer rotates head or stands still. They also require the top-view direction to be an inclined angle with less altitude for feature matching between the top and the first-person views. In our paper, we remove these requirements in our problem formulation and solution. Another series of works (Han et al., 2019, 2020a, b) try to obtain the cross-view human association and tracking, by exhaustively searching for the first-person camera and its yaw angle in the top view. In this work, we leverage the shadows to more accurately estimate both the pitch and yaw angles of first-person camera in the top view.

3 Proposed Method

3.1 Overview

In this work, we propose a simple yet effective model to relate view directions of complementary-view mobile cameras. As shown in Fig. 2, given a pair of images from synchronized first-person view and top view, respectively, we first apply a shadow detector, e.g., LISA (Wang et al., 2020) or MTMT-Net (Chen et al., 2020), to segment the shadow regions in both images. We then propose a **S**hadow **V**anishing **P**oint Detection **N**etwork (SVPN), in which the shadow maps of two views are fed into a two-stream network with different loss functions in each stream. With SVPN, we simultaneously detect the shadow vanishing point in the first-person-view image and the shadow direction in the top-view image. Finally, given the predicted vanishing point and shadow direction in the two views, respectively, we perform a geometric transformation to evaluate the view direction of the first-person view in the top view. In the following subsections, we will describe the proposed SVPN and geometric transformation in detail.

As mentioned above, shadows of different people are largely parallel in the top view, while the lines passing through them usually intersect at a common vanishing point in a first-person view after perspective projection. We unify both of them as vanishing point detection by treating the parallel shadows in the top view as intersecting at a vanishing point at infinity. In the first-person view, the shadow vanishing point may be located within or outside the image perimeter, and actually even at infinity in the degenerate cases. Considering that human shadows do not always present explicit line features, we develop a CNN-based shadow vanishing point detection network (SVPN) to locate the vanishing points in both views.

3.2 Shadow Vanishing Point Detection

First, we apply a shadow detector, e.g., Wang et al. (2020), to predict the shadow map in each view, i.e., a binary map segmenting the shadow region from the background, as shown in Fig. 3. Without detailed appearance and texture information, we only need to use a shallow network com-

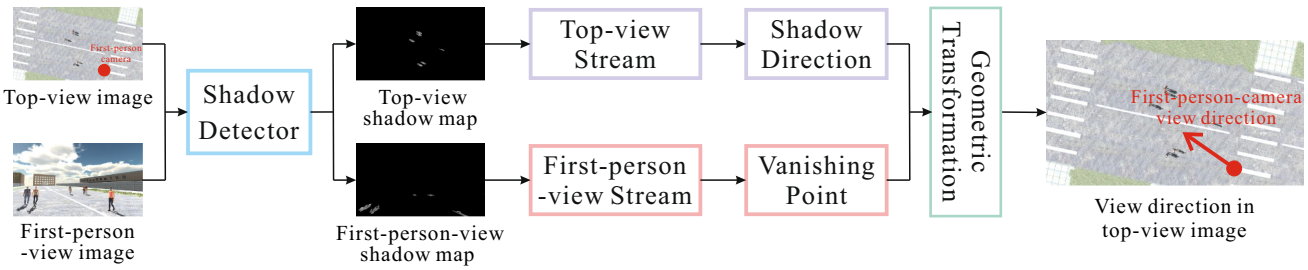


Fig. 2 Framework of the proposed complementary-view mobile camera view directions

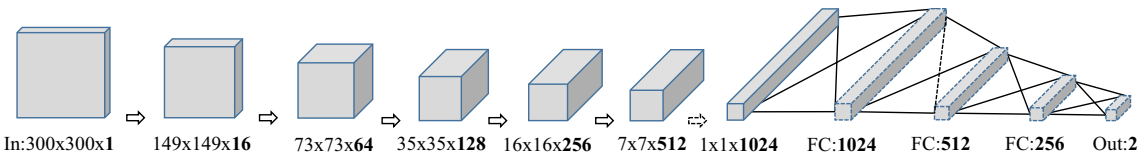
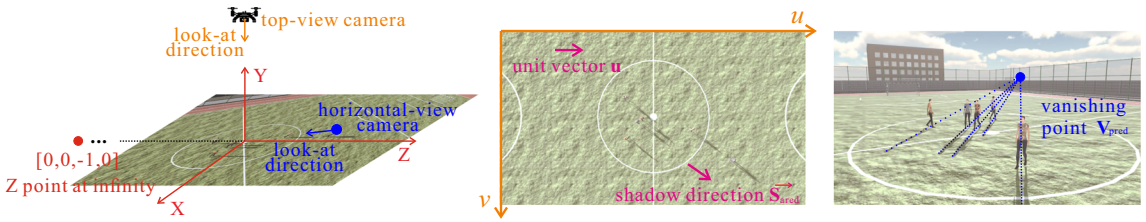


Fig. 3 An illustration of the SVPN architecture



(a) Complementary-view scene (b) Top-view image (c) Horizontal-view image

Fig. 4 An illustration of the model used for relating the top view and the first-person view in the proposed method

posed of the Convolutional Neural Network (CNN) layers to extract features from the shadow map, followed by the fully connected (FC) layers to regress the location of the vanishing point, which is shown in Fig. 3. Specifically, the SVPN network is built by two basic building blocks $\mathrm{B_C}$ and $\mathrm{B_F}$: $\mathrm{B_C}(c_1, c_2) \triangleq \{\mathrm{Conv}(3 \times 3, c_1) - \mathrm{ReLU} - \mathrm{Pool}(c_2 \times c_2)\}$, $\mathrm{B_F}(c) \triangleq \{\mathrm{FC}(c) - \mathrm{ReLU}\}$, where Conv is 2D convolution layer, and c, c_1, c_2 are parameters. The structure of the SVPN is $\{\mathrm{B_C}(16, 2) - \mathrm{B_C}(64, 2) - \mathrm{B_C}(128, 2) - \mathrm{B_C}(256, 2) - \mathrm{B_C}(512, 2) - \mathrm{B_C}(1024, 3)\}$-$\{\mathrm{B_F}(1024) - \mathrm{B_F}(512) - \mathrm{B_F}(256) - \mathrm{FC}(2)\}$. Both the top and first-person views share the same network architecture but using different loss functions.

Loss Function Given the shadow direction $\overrightarrow{\mathbf{S}_{\mathrm{pred}}} \in \mathbb{R}^2$ predicted in the top view and its corresponding ground truth $\overrightarrow{\mathbf{S}_{\mathrm{gt}}} \in \mathbb{R}^2$, we use a cosine distance loss function

$$\mathcal{L}_{\mathrm{top}} = \frac{\overrightarrow{\mathbf{S}_{\mathrm{pred}}} \cdot \overrightarrow{\mathbf{S}_{\mathrm{gt}}}}{\| \overrightarrow{\mathbf{S}_{\mathrm{pred}}} \| \| \overrightarrow{\mathbf{S}_{\mathrm{gt}}} \|}. \tag{1}$$

In the first-person view, given the predicted vanishing point $\mathbf{V}_{\mathrm{pred}} = (x_v^{\mathrm{pred}}, y_v^{\mathrm{pred}})$ and its ground truth $\mathbf{V}_{\mathrm{gt}} = (x_v^{\mathrm{gt}}, y_v^{\mathrm{gt}})$, a straightforward method for loss estimation is to calculate the distance between $\mathbf{V}_{\mathrm{pred}}$ and $\mathbf{V}_{\mathrm{gt}}$. However, each coordinate of the vanishing point may take values in the range of $(-\infty, \infty)$. When the vanishing point far away from the image center, a small error of the view direction estimation may generate very large distance between $\mathbf{V}_{\mathrm{pred}}$ and $\mathbf{V}_{\mathrm{gt}}$. This way, we use the angle error instead of the point distance to help reduce the sensibility of loss. Specifically, we first calculate the following two auxiliary angles

$$\begin{aligned} \theta_x^{\mathrm{pred}} &= \arctan \frac{x_v^{\mathrm{pred}} - x_v^{\mathrm{ctr}}}{f_x}, \\ \theta_y^{\mathrm{pred}} &= \arctan \frac{y_v^{\mathrm{pred}} - y_v^{\mathrm{ctr}}}{f_y}, \end{aligned} \tag{2}$$

where $\mathbf{O} = (x_v^{\mathrm{ctr}}, y_v^{\mathrm{ctr}})$ is the center of the image, and f_x, f_y are two focal lengths. We indicate a line between the (predicted) vanishing point and the camera optical center, and another line between the image center point and camera optical center (i.e., center axis of camera that is perpendicular to the image plane). Actually, θ_x^{pred} and θ_y^{pred} denote the angle between thees two lines along the x-axis and y-axis, respectively. Note that, $\theta_{x,y}^{\mathrm{pred}}$ and $\theta_{x,y}^{\mathrm{pred}}$ are normalized into the range of $[-\frac{1}{2}\pi, \frac{1}{2}\pi]$. Similarly, we can get θ_x^{gt} and θ_y^{gt} by replacing x_v^{pred} and y_v^{pred} with x_v^{gt} and y_v^{gt}, in Eq. (2), respectively. Then, we use Mean Square Error (MSE) function to define the loss

$$\mathcal{L}_{\mathrm{fp}} = \frac{1}{2} \sum_{i \in \{x,y\}} (\theta_i^{\mathrm{pred}} - \theta_i^{\mathrm{gt}})^2. \tag{3}$$

3.3 View Direction Relating

As shown in Fig. 4a, we assume the view direction of top-view camera is perpendicular to the ground. We set up the *world coordinate system* on the ground with the origin at the intersection point of the view direction of top-view camera and the ground plane. The projection of the sunlight onto the ground is taken as the positive direction of the Z-axis. Following the right-hand rule, we can get the X-axis and XOZ is the ground plane. In this section, for convenience *the light direction refers specifically to the direction of sunlight projection on the ground plane*.

Considering the challenge caused by the large view difference, we use light direction as an intermediate representation to relate the first-person view and the top view, by solving two sub-tasks: 1) Finding the included Euler angle $\mathbf{A} := \langle 0, 0, \mathrm{roll}^{\mathrm{top}} \rangle$ between the orientation of the top-view camera and the light direction, 2) finding the included Euler angle $\mathbf{B} := \langle \mathrm{pitch}^{\mathrm{fp}}, \mathrm{yaw}^{\mathrm{fp}}, 0 \rangle$ between the view direction of the first-person-view camera and the light direction. In this work, for the first-person camera, we only consider the yaw angle that describes the rotation parallel to the ground, and the pitch angle that reflects the look-up or head-down, since yaw and pitch are the most common transforms of the first-view wearable cameras.

For sub-task 1), as shown in Fig. 4b, we can directly obtain the relation between the light direction and the orientation of

the top-view camera by calculating the angle between the unit vector $\overrightarrow{\mathbf{u}}$, along the positive horizontal direction of the top-view image (u-axis), and the predicted shadow direction $\overrightarrow{\mathbf{S}_{\text{pred}}}$ in the pixel coordinate system of the top-view image, i.e.,

$$\text{roll}^{\text{top}} = \arccos\left(\frac{\overrightarrow{\mathbf{S}_{\text{pred}}} \cdot \overrightarrow{\mathbf{u}}}{\| \overrightarrow{\mathbf{S}_{\text{pred}}} \| \| \overrightarrow{\mathbf{u}} \|}\right). \tag{4}$$

For sub-task 2), the vanishing point in the first-person view corresponds to a point at infinity, i.e., the intersection of parallel lines, in the real world. Therefore, the shadow vanishing point in the first-person imaging plane can reflect the light direction in the real world. Combining the coordinate system defined in Fig. 4a, we establish the relation between the point at infinity on Z-axis, i.e., $\mathbf{Z}_\infty$, and the shadow vanishing point. Note that, two opposite light directions will produce the same vanishing point. In this section, we only discuss the situation as shown in Fig. 4a and we will show how to address the non-unique-solution problem in Sect. 3.4. Specifically, given the homogeneous-coordinate predicted vanishing point $\mathbf{V}_{\text{pred}} = (x_v^{\text{pred}}, y_v^{\text{pred}}, 1)^\top$ in the first-person-view image, we have

$$z\mathbf{V}_{\text{pred}} = \mathbf{K}^{\text{fp}}\,[\mathbf{R}|\mathbf{t}]\,\mathbf{Z}_\infty, \tag{5}$$

where z is a scalar factor, $\mathbf{K}^{\text{fp}}$ is the intrinsic matrix of the first-person camera, rotation matrix $\mathbf{R} \in \text{SO}(3)$ and translation vector $\mathbf{t} \in \mathbb{R}^3$ are the extrinsic parameters of the first-person camera, and $\mathbf{Z}_\infty = (0, 0, -1, 0)^\top$. Because the rotation $\mathbf{R}$ can be represented as an equivalent form of

$$\mathbf{R} = \mathbf{R}_\theta(\text{pitch}^{\text{fp}})\mathbf{R}_\psi(\text{yaw}^{\text{fp}})\mathbf{R}_\phi(0), \tag{6}$$

where $\mathbf{R}_\theta(\text{pitch}^{\text{fp}})$, $\mathbf{R}_\psi(\text{yaw}^{\text{fp}})$, $\mathbf{R}_\phi(0)$ are

$$\begin{aligned}
\mathbf{R}_\theta(\text{pitch}^{\text{fp}}) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\text{pitch}^{\text{fp}}) & -\sin(\text{pitch}^{\text{fp}}) \\ 0 & \sin(\text{pitch}^{\text{fp}}) & \cos(\text{pitch}^{\text{fp}}) \end{bmatrix}, \\
\mathbf{R}_\psi(\text{yaw}^{\text{fp}}) &= \begin{bmatrix} \cos(\text{yaw}^{\text{fp}}) & 0 & -\sin(\text{yaw}^{\text{fp}}) \\ 0 & 1 & 0 \\ \sin(\text{yaw}^{\text{fp}}) & 0 & \cos(\text{yaw}^{\text{fp}}) \end{bmatrix}, \\
\mathbf{R}_\phi(0) &= \mathbf{I}_3.
\end{aligned} \tag{7}$$

By combining the Eqs. (5)–(7), we have

$$z(\mathbf{K}^{\text{fp}})^{-1}\begin{bmatrix} x_v^{\text{pred}} \\ y_v^{\text{pred}} \\ 1 \end{bmatrix} = \begin{bmatrix} \sin(\text{yaw}^{\text{fp}}) \\ -\sin(\text{pitch}^{\text{fp}})\cos(\text{yaw}^{\text{fp}}) \\ -\cos(\text{pitch}^{\text{fp}})\cos(\text{pitch}^{\text{fp}}) \end{bmatrix}. \tag{8}$$

Hence, we can obtain the yaw and pitch components of the Euler angle $\mathbf{B}$ as

$$\begin{aligned}
\text{pitch}^{\text{fp}} &= \arctan\left(\frac{y_v^{\text{pred}} - c_y}{f_y}\right), \\
\text{yaw}^{\text{fp}} &= \arctan\left(\frac{(c_x - x_v^{\text{pred}})\cos(\text{pitch}^{\text{fp}})}{f_x}\right),
\end{aligned} \tag{9}$$

where f_x, f_y are two focal lengths and c_x, c_y are the coordinates of the first-person camera's principal point, in the intrinsic matrix $\mathbf{K}^{\text{fp}}$. Note that, here we use the pinhole camera model for convenience.

After getting the Euler angles $\mathbf{A}$ and $\mathbf{B}$, we can map the view direction of the first-person camera to the top view by

$$\text{yaw} = \text{yaw}^{\text{fp}} + \text{roll}^{\text{top}}, \quad \text{pitch} = \text{pitch}^{\text{fp}}, \tag{10}$$

where yaw angle is between the view direction of the first-person-view camera and the u-axis of the pixel coordinate system of the top-view image, and pitch angle is between the view direction of first-person-view camera and the imaging plane of the top-view camera.

3.4 Implementation Details

Network Training As shown in Fig. 2, to train the proposed SVPN, we use the human shadow maps as input, as well as the corresponding $\overrightarrow{\mathbf{S}_{\text{gt}}}$ and $\mathbf{V}_{\text{gt}}$ as output. Note that, in real-world scenes, the ground-truth (human) shadow maps and vanishing points are very difficult to obtain. This way, we develop a controllable shadow generation tool ShadowX to automatically get the accurate shadow maps and vanishing points without manual annotation, which will be described in detail in Sect. 4.2. We use the different shadow detectors to generate the shadow map for SVPN. Specifically, during training, we first feed the ground-truth shadow segmentation maps, which are automatically generated by our ShadowX, into the SVPN for pre-training. After that, we fine tune the pre-trained model by feeding the shadow maps generated by a state-of-the-art shadow detection algorithm, i.e., LISA (Wang et al., 2020). In the testing stage, we use the LISA algorithm to generate shadow maps.

Strategy for Ambiguity Elimination In Sect. 3.3, we have mentioned that a vanishing point (in the first-person view) may be generated from two opposite candidate light directions, from which we have to ensure the unique light direction. Therefore, we use a simple and effective strategy based on the prior observation that the subject itself is always located between the light and its shadow (in both the top- and first-person view images). This way, in the first-person view, we estimate the rough light location by the above prior

observation and further select the unique light direction from two candidates. Similarly, we should also ensure the unique shadow direction in the top-view image, where we directly feed the instance shadow detection results with the locations of both the subjects and shadows to regress the shadow direction using the proposed top-view stream of SVPN, which can be regarded as implicitly including the above prior observation by taking the relative position of the subjects and shadows as input.

4 Proposed Dataset

We do not find publicly available datasets with top-view and first-person-view images with ground-truth annotations of shadow direction and camera directions. Actually, without additional measuring devices, it is very difficult to get the ground-truth view directions for such complementary views. The vanishing point in the first-person view is also hard to annotate accurately, especially when it is located far away from the image center. Therefore, we instead develop a controllable engine to generate a new synthetic dataset to ensure the reliability and accuracy of the annotations, which also facilitate more in-depth analysis.

4.1 System Configuration

Controllable Shadow Generation Tool—ShadowX We build the tool namely ShadowX for simulation data generation, which is built on Unity 3D (Riccitiello, 2018). Specifically, we create a 3D controllable world with changeable person models and scenes. The person models are from PersonX (Sun & Zheng, 2019) project and the scenes are from Unity Store. The virtual world is controllable on person position, camera position, illumination and shadow rendering. Thus, we can use ShadowX to build various and computable environments and generate corresponding images with accurate annotations.

All these configurations in ShadowX are editable, which are described as below.

• Illumination: Illumination parameters in ShadowX can be set freely. Light source can be directional light, spot light, point light, etc. Other parameters, including the direction, position, number, and intensity of lights, can be customized and modified. Different combinations of these them make shadow generation controllable and computable.

• Cameras: The intrinsic parameters of cameras, such as field of view, resolutions, and focal length, and the extrinsic parameters of cameras, such as transformation and rotations, can also be accessed and modified in ShadowX.

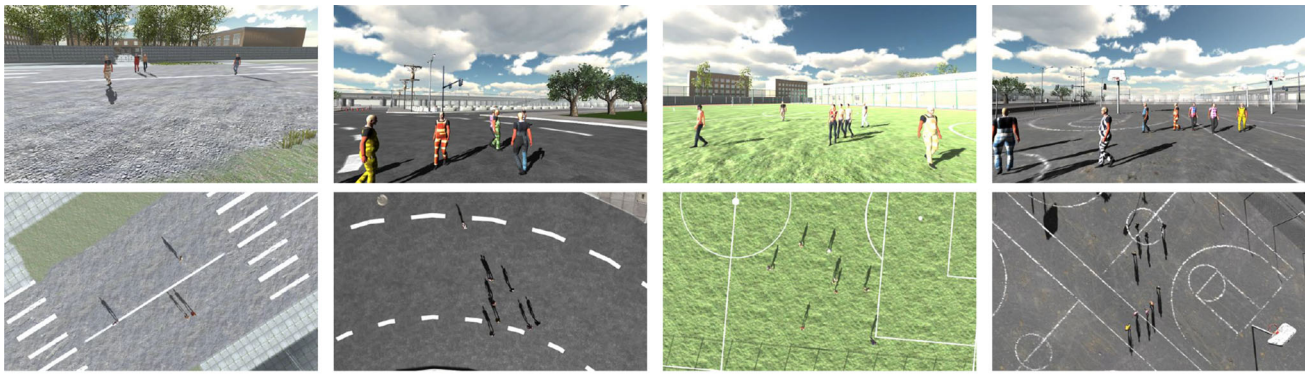
• Subjects: The settings of subjects in each scene are controllable, including the number of subjects, the initial position of each person, and the moving direction of each person at each time.

4.2 Synthetic Dataset

Specifications for Dataset Construction With ShadowX, we generate a new synthetic dataset for model training and quantitative performance evaluation of our method. We select several lifelike virtual environments, e.g., the basketball court, soccer field, city street and campus as the backgrounds in our dataset and different backgrounds have impact on the shadow detection. We use human models from PersonX (Sun & Zheng, 2019) as the people in each scene. In each image, there are 2 - 10 persons walking in the scene, and one of them wears a wearable camera overhead to observe the other people. Six factors are set to be random to ensure the variety and diversity of the synthesized images: (i) the position of the first-person-view camera, (ii) the pitch angle and yaw angle of the first-person-view camera, (iii) the roll angle of the top-view camera, (iv) the light direction, (v) the position of each person, (vi) the moving direction of each person. We employ a single point light source at infinity to simulate the sunlight. The range of the pitch and yaw angle of the first-person-view camera are set as 30° and 360°, respectively. The range of the roll angle of the top-view camera is 360°. Note that, all the factors are random but with rough balance in terms of the number of synthesized images, to prevent from undesired biases to certain factors. Some sample images are shown in Fig. 5.

Dataset Statistics We generate in total 8000 images, i.e., 4000 pairs of the top- and first-person views, of the resolution of 968 × 545 for our dataset. We use ShadowX to generate rich annotations including 3D coordinates the persons and cameras, segmentations of human shadows, and Euler angles between the light and the top/first-person view directions. We further perform geometric transform to get 2D pixel coordinates of person and camera locations and the vanishing points. We split the dataset into training data and testing data by ratio 4:1 with no overlap.

Training Data Generation *Shadow map generation:* For the shadow maps, we first generate the images with shadows and then create the images without the shadows by closing the shadow rendering. This way, we can obtain the shadow map through the difference between the images with and without shadows. *Vanishing point generation:* We can also get the ground-truth vanishing points in ShadowX. In specific, with ShadowX, given the sunlight direction and the human location in the world coordinate system, we can get the 3D human shadow vectors of all the subjects on the ground (in the world coordinate system), which are mutual parallel. Also, with the intrinsic parameters and extrinsic parame-



(a) Campus (b) City Street (c) Football Ground (d) Basketball Court

Fig. 5 Sample image pairs in the proposed synthetic dataset. Top: first-person view. Bottom: Top view

ters of the cameras, we can transform the shadow vectors into the top-view and first-person-view images, respectively. In the top view, the 2D human shadow vectors keep parallel to each other. Then we can get the shadow direction $\overrightarrow{\mathbf{S}_{\text{gt}}}$ defined in Eq. (1). In the first-person view, we can also get all the 2D human shadow vectors. Considering the properties of perspective transformation, all these shadow vectors will intersect into the vanishing point. This way, we get the location of the vanishing point on the first-person-view image, i.e., $\mathbf{V}_{\text{gt}}$, by calculating the intersection of the 2D shadow vectors.

4.3 Evaluation Metrics

We use MAE (Mean Absolute Error) score as the evaluation metric in three tasks, i.e., **Task I** of first-person-view vanishing point prediction, **Task II** of top-view shadow direction prediction, and **Task III** of relating the first-person view and the top view. MAE is computed as $\mathcal{M} = \frac{\sum_{i=1}^{n} |\Delta(\hat{y}_i, y_i)|}{n}$, where $\hat{y}_i$ and y_i are the predicted and ground-truth values of the i-th image. In different tasks, Δ is defined differently, including Δ_V, Δ_S, Δ_P, Δ_Y as described below.

For **Task I**, on the first-person view, let $\mathbf{V}_{\text{pred}} = (x_v^{\text{pred}}, y_v^{\text{pred}}, 1)$ and $\mathbf{V}_{\text{gt}} = (x_v^{\text{gt}}, y_v^{\text{gt}}, 1)$ be the homogeneous coordinates of the predicted vanishing point and the ground-truth vanishing point, respectively, and $\mathbf{O} = (x_v^{\text{ctr}}, y_v^{\text{ctr}}, 1)$ be the center of the image. $\mathbf{C}_{\text{fp}}$ is the optical center of the first-person-view camera. The coordinate-based error between the predicted and the ground-truth vanishing points are

$$\begin{aligned} \overrightarrow{\mathbf{C}_{\text{fp}}\mathbf{V}_{\text{pred}}} &= (x_v^{\text{pred}} - x_v^{\text{ctr}}, y_v^{\text{pred}} - y_v^{\text{ctr}}, f), \\ \overrightarrow{\mathbf{C}_{\text{fp}}\mathbf{V}_{\text{gt}}} &= (x_v^{\text{gt}} - x_v^{\text{ctr}}, y_v^{\text{gt}} - y_v^{\text{ctr}}, f), \end{aligned} \tag{11}$$

where f is the camera focal length. As discussed above, we combine them into an angle-based error metric by

$$\begin{aligned} &\Delta_V(\mathbf{V}_{\text{pred}}, \mathbf{V}_{\text{gt}}) \\ &= \arccos\left(\frac{\overrightarrow{\mathbf{C}_{\text{fp}}\mathbf{V}_{\text{pred}}} \cdot \overrightarrow{\mathbf{C}_{\text{fp}}\mathbf{V}_{\text{gt}}}}{\| \overrightarrow{\mathbf{C}_{\text{fp}}\mathbf{V}_{\text{pred}}} \| \| \overrightarrow{\mathbf{C}_{\text{fp}}\mathbf{V}_{\text{pred}}} \|}\right), \end{aligned} \tag{12}$$

For **Task II**, in the top view, let $\overrightarrow{\mathbf{S}_{\text{pred}}}$ and $\overrightarrow{\mathbf{S}_{\text{gt}}}$ be the predicted and ground-truth (normalized) direction vectors of the shadows. We define the included angle between $\overrightarrow{\mathbf{S}_{\text{pred}}}$ and $\overrightarrow{\mathbf{S}_{\text{gt}}}$ as the error by

$$\Delta_S(\overrightarrow{\mathbf{S}_{\text{pred}}}, \overrightarrow{\mathbf{S}_{\text{gt}}}) = \arccos\left(\frac{\overrightarrow{\mathbf{S}_{\text{pred}}} \cdot \overrightarrow{\mathbf{S}_{\text{gt}}}}{\| \overrightarrow{\mathbf{S}_{\text{pred}}} \| \| \overrightarrow{\mathbf{S}_{\text{gt}}} \|}\right). \tag{13}$$

For **Task III**, we first evaluate the relative pitch and yaw angles between the two views as defined in Eq. (10), using the errors Δ_P and Δ_Y to be the difference between ground truth and the prediction. We further define δ_P and δ_Y to be a normalized version of Δ_P and Δ_Y as $\delta_P = \frac{\Delta_P}{D_P} \times 100\%$, $\delta_Y = \frac{\Delta_Y}{D_Y} \times 100\%$, where $D_P = 30°$ and $D_Y = 360°$ are range of the pitch and yaw angle, respectively, used in our data generation. Then we compute average of δ_P and δ_Y as the overall error δ_A. Besides that, we also use 10% and 20% as the threshold for δ_A for each image to count for the true predictions, and evaluate the corresponding accuracy as Acc@10 and Acc@20, respectively.

5 Experimental Results

5.1 Setup

We use Pytorch backend for implementing the proposed network and run on a computer with RTX 2080Ti GPU. Before

Table 1 Comparative results of different methods, where '–' denotes the ablative results that are not influenced, and '/' denote the ablative results that can not be obtained by the corresponding method

	$\Delta_V \downarrow(^\circ)$	$\Delta_S \downarrow(^\circ)$	$\Delta_P \downarrow(^\circ)$	$\Delta_Y \downarrow(^\circ)$	$\delta_p \downarrow(\%)$	$\delta_Y \downarrow(\%)$	$\delta_A \downarrow(\%)$	Acc@10↑	Acc@20 ↑
D2-Net + 5 points	/	/	81.43	89.56	271.43	24.88	148.16	8.48	11.13
Samp. w prior	91.39	92.49	20.23	89.00	67.43	24.72	46.08	2.38	11.26
Shadow + direct.	83.90	89.07	21.59	88.67	71.96	24.63	48.29	27.95	29.80
w Eucli. loss	70.28	–	13.20	55.88	44.01	15.52	29.77	10.86	32.05
w/o select.	–	–	–	59.77	–	16.60	11.17	61.85	72.32
w/o hum. in top	–	26.06	–	36.67	–	10.19	7.96	75.89	87.55
w/o shadow for tuning	20.94	65.31	1.88	68.00	6.26	18.89	12.57	54.17	69.54
w/o GT shadow train.	20.45	67.84	1.90	72.41	6.34	20.12	13.23	53.38	65.96
w GT shadow test.	16.19	2.78	1.61	15.13	5.38	4.20	4.79	90.34	98.24
Ours	20.34	9.18	1.72	22.85	5.73	6.35	6.04	83.84	94.70

training, we resize the image into 300 × 300. Our network is trained on 3200 image pairs for 50 epochs with the initial learning rate 0.001. The inference time of our model is over 12 fps.

5.2 Results

Baselines We do not find directly related comparative methods. Given the large view difference between the complementary-view images, most methods for multi-view camera pose estimation are not applicable here. We consider the following three baseline methods.

- D2-Net + 5 points: We apply a recent feature point detection method D2-Net (Dusmanu et al., 2019) on the first-person-view and top-view images to extract the keypoints. Then we use the classical five-point method (Nister, 2004) to estimate relative pose between two cameras to get the yaw and pitch angles.
- Sampling w prior: Supposing the view direction range of the first-person-view camera wearer is known, we randomly generate the vanishing point coordinate in the first-person-view image and shadow direction in the top view.
- Shadow + direct.: We first use Wang et al. (2020) to generate an instance shadow map, then we connect the lines between the feet of the subjects and the center of corresponding shadows, and compute the geometric center of the intersections of each line pair as the vanishing point.

Ablation Study We consider several model variants for ablation study.

• w Eucli. loss: Use Euclidean distance between the predicted and ground-truth vanishing point as the loss function instead of the one proposed in Eq. (2).

• w/o select.: In the first-person view, we randomly select the light direction from two candidate directions without the ambiguity elimination strategy using the prior as discussed in Sect. 3.4.

• w/o hum. in top: In top view, we use only shadow detection but without the human detection results as the input of the proposed shadow vanishing point detection network.

• w/o shadow for tuning: Remove the fine-tuning stage of using detected shadows for SVPN training as discussed in network training part.

• w/o GT shadow for train.: In the training stage, we use the predicted shadow map instead of the ground-truth shadows.

• w GT shadow for test.: In the testing stage, we use the ground-truth shadow map instead of the predicted shadows.

Comparison Result Analysis As shown in the top of Table 1, all the baseline methods produce very poor results in our task, which demonstrates that *our problem is non-trivial*. Traditionally, a common method for relative pose estimation is based on the matched keypoint pairs. For example, we apply an effective feature point extraction and matching methods D2-Net and the classical camera pose estimation method 5 points algorithm. The results are very poor, since given the huge view difference, it's almost impossible to find any accurately matched keypoint pairs.

Similarly, the comparative method 'Shadow + direct.' also generates a poor performance, which uses a straightforward approach for vanishing point detection using the human shadows. The comparison of it with the proposed method demonstrates the effectiveness of the proposed SVPN.

Ablative Analysis From Table 1, we can see that: (1) our model using Euclidean distance loss has a poor performance, which reflects the effectiveness of the proposed loss defined in Eq. (2). (2) 'w/o select.' performs worse when relating the first-person view and the top view, e.g., the performance

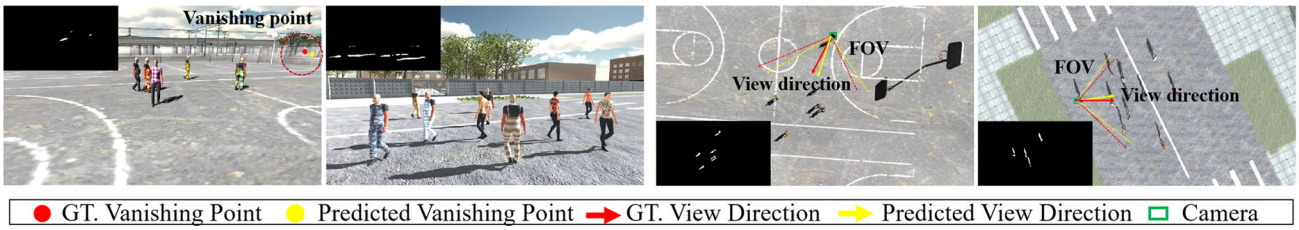


Fig. 6 Qualitative results from the synthetic dataset

using the metric Δ_Y, which denotes the necessity of the strategy for ambiguity elimination in the first-person view. (3) 'w/o hum. in top' also provides a relatively poor result. This indicates that the joint use of both human detection and shadow detection results in top view as input (as discussed in Sect. 3.4) works better than only using shadow. This is because the input of both human detection and shadow detection makes use of the relative position of the subjects and shadows, which can help to estimate the direction of the shadow vector in the top view. (4) The accuracy of shadow detection can not be guaranteed, especially in the top view, which makes the predicted shadow to be of large difference from the ground truth. As a result, performance gets worse when excluding either the proposed predicted shadow fine-tuning or ground-truth training. Therefore, we train our model on ground-truth shadow maps and fine-tune on predicted ones to promote the generalization ability of our model. (5) The last row shows the results using ground-truth shadow maps in testing stage, which naturally represents the best performance.

Overall, from Table 1 we can see that, the proposed method outperforms the existing comparison methods by a large margin, which demonstrates the superiority of the proposed method. We can also see from Table 1 that the main components in our method are effective.

5.3 Qualitative Analysis

Figure 6 shows the qualitative results on two cases in our dataset. We can see that all the shown cases have an accurate view-direction estimation results, where the errors are within 10 degrees. Note that, our method also works when other non-person shadows are present, e.g., the shadow of the basketball stands in Fig. 6. Actually, the proposed method first detects shadows and then use the shadow map as the input of the following components and different image style and background (even across the synthetic and real-world scenes) only have impact on the shadow detection.

5.4 Applied Condition Analysis

We build the flexible configuration for simulation data generation, which makes the scene conditions controllable and various. This way, we can conduct the detailed experimental analysis under different applied conditions. We then investigate the performance of the proposed method under different condition including the number of subjects, light direction, angle between view and light directions and shadow length, respectively.

Fig. 7 Condition analysis of the number of subjects

Fig. 8 Condition analysis of the light direction

Fig. 9 Condition analysis of the angle between view and light directions

Fig. 10 Condition analysis of the shadow length

Number of Subjects As shown in Fig. 7a, the performance of our method gets worse when the number of subjects in the scene is too large or too small. It can be explained that when the scene is too crowded as in Fig. 7b, the mutual occlusions in the first-person view may prevent the accurate shadow detection. When the subject number is too small as in Fig. 7c, there are insufficient instances of shadows for accurate shadow direction prediction.

Fig. 11 Examples from the real-world images. The cases are tagged with various labels according the attributes of the scene, including Clear shadow (CS), Broken shadow (BS), Disordered shadow (DS), Complex background (CB), Open and clear area (OA), Uneven ground (UG), Dense crowd (DC), Irregular shape (IS), Summer (SU), Winter (WI)

Light Direction We vary the light angle with the ground in the data generation and evaluate its effect on the final results. Two examples are shown in Fig. 8b, c—with a small light angle, shadows are longer but the image is darker, while with a large angle, shadows are shorter but the image is brighter. The curve shown in Fig. 8b shows that with increased light angle, the shadow region has better contrast, leading to more accurate shadow detections and better performance in relating two views.

Angle between view and light directions As shown in Fig. 9a, the angle between the first-person view direction and the light direction also influences the results. When this angle approaches 90° as shown in Fig. 9b, shadows are parallel to the imaging plane and the shadow vanishing point is at infinity. In this case, the MAE reaches the highest. When this angle is further away from 90°, as shown in Fig. 9c, the shadow vanishing point can be detected with higher accuracy, leading to lower MAEs.

Shadow Length in Image The effects of the shadow length on the final results are shown in Fig. 10. We can see that either overly long or overly short shadows will hurt the final accuracy. A possible reason is that overly short shadows are prone to be missed in shadow detection, as shown in Fig. 10b, while overly long shadows introduce more mutual occlusions or overlaps, as shown in Fig. 10c.

From the above analysis, we can see that, although with minor performance fluctuations, our method is effective under different conditions, which shows the robustness of the proposed method.

5.5 Real-World Case Analysis

We further apply the proposed method for the real-world practical scenarios with various factors. We consider the factors about the quality of the human shadow, background of scene (color and flatness of the ground, clutters on the ground), density of crowd, season and weather, etc. We provide different labels according to these factors, which are assigned to each real-world case as shown in Fig. 11. The labels are defined as below.

Clear shadow (CS): The shadows are easy to be detected and the predicted shadow maps are clear.

Broken shadow (BS): The predicted shadow maps are broken since the mutual occlusions in the image or the prediction errors of the shadow detection algorithm.

Disordered shadow (DS): The human shadows are disordered by the shadows of the other objects appearing in the scene, e.g., trees, buildings.

Complex background (CB): The background is complex with various objects on the ground.

Open and clear area (OA): The scene is open without many clutters.

Uneven ground (UG): The ground is uneven with ups and downs.

Dense crowd (DC): The crowd on the ground is relatively dense.

Irregular shape (IS): The human shape is irregular, e.g, with an umbrella, bag, etc. This makes the human shadow is also irregular.

Summer (SU)/Winter (WI): The images are taken in the summer time or winter time.

Figure 11 shows the results on several pairs of real images by applying the proposed method. Note that, we can not conduct the quantitative evaluation since the accurate ground-truth results for our problem are very difficult to obtain. We have tried the gyroscopes integrated in the smart phones and cameras, which, however, fail to solve the proposed problem because of the random drift error produced by the external disturbances. This way, we manually annotate the shadow vanishing point in the first-person view, and the view direction of the first-person view camera in the top view. The cases in Fig. 11 are tagged with several labels according to the attributes of the scene.

Specifically, we can see from Fig. 11a, b that the human shadows are clear in both views, which makes the predicted shadow maps are also with high quality. In this case, the proposed method provides an accurate view-direction estimation result, where the errors are within 10 degrees. From Fig. 11c, d, we can see that the predicted shadow map is not very perfect. This is because the shadow of the tree branches in the top view in (c) and the occlusions in the first-person view in (d) make the predicted human shadow map broken. Even so, the proposed method also provides an acceptable view direction estimation result. At the third row in Fig. 11, we show two examples taken in an open area. Although the ground is clear, we can see that the shadows are not very strong and long since the time of day and weather when the image was shot. Also, the irregular human shape (carrying the umbrella, bag) also makes the shadows irregular. Under the circumstances, we can see that our method still provides a good performance. Note that, our method does not require the predicted shadow map to include all the human shadows in the scene. This alleviates the dependence of our method on the high accuracy of shadow prediction algorithms. At the fourth row in Fig. 11, we provide two challenging scenes where the ground is uneven with parterre or stairs. The human shadows are distorted, especially in the first-person view, influenced by the undulation of the ground. We can see that the proposed method can handle these cases and provide a promising result. Finally, in the last row, we show two cases that our method can not handle very well. As shown in Fig. 11i, the human shadows are drowned in the shadows of a tangle of branches. In Fig. 11j, the human shadows are indiscoverable given the overcast sky. In this case, the proposed method can not work well under the very poor shadow map.

Overall, the above real-world case study verifies that the proposed method can generally handle the real-world cases well. Note that, the SVPN model used here is trained only on synthetic data and directly applied to the real-world images. By applying the shadow detector trained on the real images, our method is not heavily dependent on the high accuracy of shadow prediction results. This makes the proposed method *not sensitive to complex real-world scenes* and has *good cross-domain generalization ability*.

6 Discussion

6.1 Assumption and Limitation Analysis

Camera Setting In this work, we assume the top-view camera is perpendicular to the ground and the first-person camera is (roughly) parallel to the ground with a moderate pitch angle, which is recognized in the previous works (Han et

al., 2022; Ardeshir & Borji, 2018a; Han et al., 2020b). This assumption is not strict since it is aligned to the general settings of the top view from a drone-mounted camera, that takes a global picture of all subjects on the ground, and the first-person view from a wearable camera, which aims to keep the subjects not out of its FOV.

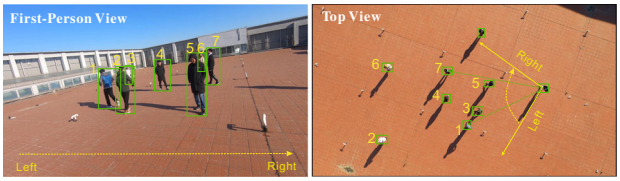
Scene Setting As discussed in Sect. 5.4, the proposed method may not work well when the scene is too crowded, e.g., with many overlapped shadows, or the light condition is poor, e.g. with unclear shadows in cloudy days.

With these limitations, this work has not addressed all complex issues in real-world applications. But we propose a brand-new approach to address this new challenging problem. We will extend our work to more complex scenes in the future.

6.2 Advantages and Applications

Advantages From the perspective of performance, the proposed method addresses a foundational problem of complementary-view camera (weak) calibration for which existing methods are not useful. For example, existing vision based methods mostly apply key-point detection and matching for camera calibration. However, key-point features usually cannot be correctly matched across the complementary views given the enormous view difference. This can be also seen from the quantitative evaluation results in Table 1. Also, the hardware based methods, e.g., Gyroscopes integrated with smart-phones and cameras, cannot be used to solve this problem either since external disturbances produce random drift error all the time. The proposed method can obtain the camera relative pose (yaw and pitch angles) estimation results *frame by frame* with a relatively high accuracy and an efficient running speed (over 12 fps), which is available for many applications. We then discuss about the potential advantage w.r.t. the practicability for several downstream tasks, including the large-view-difference camera calibration, cross-view human identification and co-attention human detection, as follows.

Large-View-Difference Camera Calibration It is a challenging problem to calibrate the external parameters of multiple cameras with large FOV (field-of-view) difference, especially in mobile camera groups composed of top- and horizontal-view cameras. The core of external parameter calibration of multiple cameras is the feature matching between the captured images. However, such a large FOV difference not only makes traditional feature extraction and matching methods e.g., SIFT (Lowe, 2004), but also the deep learning based methods, e.g., D2-Net (Dusmanu et al., 2019), fail to play a stable role in this situation. Therefore, most approaches choose to use additional optical sensors such as LIDAR, or set up virtual markers in the field of view to calibrate the cameras. We present a computer vision guided method to steadily relate the top- and horizontal-view (consumer-level) cameras *without using other sensors*. The yaw and pitch angles of the horizontal camera relative to the top camera can be used as a prior information in external parameter calibration of such two totally different cameras. For visual application systems, e.g., SfM (Structure from Motion), which needs frequent calibration of external parameters between cameras, the stable estimation of yaw and pitch angles between top and horizontal cameras can (1) limit the search space of feature matching, and reduce the interference of large view angle difference on feature matching, and (2) reduce the original six DOF (degrees of freedom) into four, thus reducing the complexity of external parameter estimation.



(a) (b)

Fig. 12 An illustration of the cross-complementary-view human identification

Cross-View Human Identification Complementary-view human identification, i.e., identifying the same persons across the top and first-person view, is a *fundamental problem* and has many applications. For example, in an outdoor scenario without pre-installed cameras, we can associate the humans taken by the cameras on a drone (top view) and worn by several law enforcement officials (first-person views) for collaborative tracking, localization, and individual/group activity recognition, etc. This is also a *very challenging problem* due to the large view difference between such two views, which makes the appearance and motion features to show huge difference (Han et al., 2022). If we can obtain the view directions, i.e., the yaw and pitch angles of the first-person-view camera, we can match the persons across these two views by examining their spatial layout. Specifically, as shown in Fig. 12a, in the first-person view, we arrange the persons appearing in the image from left to right. Correspondingly, with the estimated *yaw angle* and the FOV (field of view) of the first-person camera in top view, we orderly arrange the persons locating at the rays starting from the left to right boundary of the FOV, as shown in Fig. 12b. This way, we can (coarsely) match the subjects across two views, which can be used as a constraint together with other features for the cross-view human identification. For such cross-view association, we can further consider the relation between the subjects' height (in terms of bounding box) in the first-person view and the subjects' depth (relative to the first-person-view

Fig. 13 An illustration of the co-attention human detection using complementary-view cameras with the relative view direction

camera) in the top view, as detailedly discussed in Han et al. (2022). However, such algorithm assumes the first-person view is horizontal to the ground, which is not always true in practice. This assumption can be removed by combining with the *pitch angle* of the first-person view camera estimated in this work.

Co-attention Human Detection Important person detection is a significant task in video surveillance. Previous works have studied to localize the important persons with the help of surrounding people's visual attention (Fan et al., 2018, 2019; Chong et al., 2018; Recasens et al., 2015; Han et al., 2020b). Specifically, as shown in Fig. 13, by relating the view directions of complementary-view cameras, we can map the view directions of multiple first-person cameras to the unified global top view, in which we can estimate the visual-attention regions of each first-person camera. Based on the visual attention of each camera wearer and the co-attention fusion strategy, e.g., the one proposed in Han et al. (2020b), we can identify the person that draws the attention of most people at any time. Figures 12 and 13 show two downstream tasks based on the proposed complementary-view camera relating problem. This demonstrates that the problem in this work is a fundamental problem, which can be regarded as the first step to connect the complementary-views and support their collaborative video analysis, especially the crowd analysis.

7 Conclusion

In this work, we have studied a new problem of relating the view directions of complementary first-person and top views by leveraging human shadows. We proposed a new shadow vanishing point detection network to simultaneously get the shadow vanishing point in the first-person-view image and the shadow direction in the top-view image, based on which, we established a geometric transformation to estimate the pitch and yaw angles between the two camera views. As a weaker version of relative camera pose calibration, the derived relative pitch and yaw angles can be used to promote many important applications. We used the controllable shadow generation engine ShadowX, to construct a synthetic dataset for quantitative performance evaluation. The experimental results verified the effectiveness of the proposed method.

Acknowledgements This work was supported in part by the NSFC under Grants U1803264, 62072334.

References

Antunes, M., & Barreto, J. P. (2013). A global approach for the detection of vanishing points and mutually orthogonal vanishing directions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'13* (pp. 1336–1343).

Ardeshir, S., & Borji, A. (2016). Ego2top: Matching viewers in egocentric and top-view videos. In *Proceedings of the European Conference on Computer Vision, ECCV'16* (pp. 253–268).

Ardeshir, S., & Borji, A. (2018a). Egocentric meets top-view. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(6), 1353–1366.

Ardeshir, S., & Borji, A. (2018b). Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *Proceedings of the European Conference on Computer Vision, ECCV'18* (pp. 285–300).

Ardeshir, S., Regmi, K., & Borji, A. (2016). Egotransfer: Transferring motion across egocentric and exocentric domains using deep neural networks. arXiv:1612.05836.

Balcı, H., & Güdükbay, U. (2017). Sun position estimation and tracking for virtual object placement in time-lapse videos. *Signal, Image and Video Processing, 11*(5), 817–824.

Barekatain, M., Martí, M., Shih, H. F., Murray, S., & Prendinger, H. (2017). Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, CVPRW'17*.

Barinova, O., Lempitsky, V., Tretiak, E., & Kohli, P. (2010). Geometric image parsing in man-made environments. In *Proceedings of the European Conference on Computer Vision, ECCV'10* (pp. 57–70).

Barnard, S. T. (1983). Interpreting perspective images. *Artificial Intelligence, 21*(4), 435–462.

Birdal, T., Bala, E., Eren, T., & Ilic, S. (2016). Online inspection of 3D parts via a locally overlapping camera network. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV'16* (pp. 1–10).

Bolles, R. C., & Fischler, M. A. (1981). A RANSAC-based approach to model fitting and its application to finding cylinders in range data. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI'81* (pp. 637–643).

Borji, A. (2016). Vanishing point detection with convolutional neural networks. arXiv:1609.00967.

Censi, A., Franchi, A., Marchionni, L., & Oriolo, G. (2013). Simultaneous calibration of odometry and sensor parameters for mobile robots. *IEEE Transactions on Robotics, 29*(2), 475–492.

Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., & Heng, P. A. (2020). A multi-task mean teacher for semi-supervised shadow detection. In (Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'20) (pp. 5611–5620).

Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., & Rehg, J. M. (2018). Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision, ECCV'18* (pp. 383–398).

Coughlan, J. M., & Yuille, A. L. (1999). Manhattan world: Compass direction from a single image by bayesian inference. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV'99* (pp. 941–947).

Doğan, Y., Sonlu, S., & Güdükbay, U. (2021). An augmented crowd simulation system using automatic determination of navigable areas. *Computers & Graphics, 95*, 141–155.

Dong, S., Shao, X., Kang, X., Yang, F., & He, X. (2016). Extrinsic calibration of a non-overlapping camera network based on close-range photogrammetry. *Applied Optics, 55*(23), 6363–6370.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., & Sattler, T. (2019). D2-Net: a trainable CNN for joint description and detection of local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'19* (pp. 8092–8101).

Fan, L., Chen, Y., Wei, P., Wang, W., & Zhu, S. C. (2018). Inferring shared attention in social scene videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'18* (pp. 6460–6468).

Fan, L., Wang, W., Huang, S., Tang, X., & Zhu, S. C. (2019). Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV'19* (pp. 5724–5733).

Guan, B., Zhao, J., Li, Z., Sun, F., & Fraundorfer, F. (2021). Relative pose estimation with a single affine correspondence. *IEEE Transactions on Cybernetics*. https://doi.org/10.1109/TCYB.2021.3069806

Han, R., Zhang, Y., Feng, W., Gong, C., Zhang, X., Zhao, J., Wan, L., & Wang, S. (2019). Multiple human association between top and horizontal views by matching subjects' spatial distributions. arXiv:1907.11458.

Han, R., Feng, W., Zhao, J., Niu, Z., Zhang, Y., Wan, L., & Wang, S. (2020a). Complementary-view multiple human tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI'20* (pp. 10917–10924).

Han, R., Zhao, J., Feng, W., Gan, Y., Wan, L., & Wang, S. (2020b). Complementary-view co-interest person detection. In *Proceedings of the ACM International Conference on Multimedia, ACM MM'20* (pp. 2746–2754).

Han, R., Feng, W., Zhang, Y., Zhao, J., & Wang, S. (2022). Multiple human association and tracking from egocentric and complementary top views. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 5225–5242.

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.

Hough, P. V. (1959). Machine analysis of bubble chamber pictures. In *Proceedings of the International Conference on High Energy Accelerators and Instrumentation* (pp. 554–556).

Kluger, F., Ackermann, H., Yang, MY., & Rosenhahn, B. (2017). Deep learning for vanishing point detection using an inverse gnomonic projection. In *German Conference on Pattern Recognition* (pp. 17–28).

Kogecka, J., & Zhang, W. (2002). Efficient computation of vanishing points. In *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'20* (pp. 223–228).

Lee, S., Kim, J., Yoon, J. S., Shin, S., Bailo, O., Kim, N., Lee, T. H., Hong, H. S., Han, S. H., & Kweon, I. S. (2017). VPGNet: Vanishing point guided network for lane and road marking detection and recognition. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV'17* (pp. 1947–1955).

Lezama, J., Grompone von Gioi, R., Randall, G., & Morel, J. (2014). Finding vanishing points via point alignments in image primal and dual domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'14* (pp. 509–515).

Li, T., Liu, J., Zhang, W., Ni, Y., & Li, Z. (2021). UAV-Human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'21* (pp. 16266–16275).

Lin, Y., Ezzeldeen, K., Zhou, Y., Fan, X., Yu, H., Qian, H., & Wang, S. (2015). Co-interest person detection from multiple wearable camera videos. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV'15* (pp. 4426–4434).

Liu, P., Yang, P., Wang, C., Huang, K., & Tan, T. (2016). A semi-supervised method for surveillance-based visual location recognition. *IEEE Transactions on Cybernetics, 47*(11), 3719–3732.

Liu, Z., Li, F., & Zhang, G. (2014). An external parameter calibration method for multiple cameras based on laser rangefinder. *Measurement, 47*, 954–962.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.

Magee, M. J., & Aggarwal, J. K. (1984). Determining vanishing points from perspective images. *Computer Vision, Graphics, and Image Processing, 26*(2), 256–267.

Micusik, B. (2011). Relative pose problem for non-overlapping surveillance cameras with known gravity vector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'11* (pp. 3105–3112).

Miraldo, P., Araujo, H., & Goncalves, N. (2015). Pose estimation for general cameras using lines. *IEEE Transactions on Cybernetics, 45*(10), 2156–2164.

Nister, D. (2004). An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(6), 756–777.

Perera, A. G., Law, Y. W., & Chahl, J. (2019). UAV-gesture: A dataset for UAV control and gesture recognition. In *Proceedings of the European Conference on Computer Vision Workshop, ECCVW'19*.

Recasens, A., Khosla, A., Vondrick, C., & Torralba, A. (2015). Where are they looking? In *Proceedings of the Advances in neural information processing systems, NeurIPS'15* (vol. 28).

Riccitiello, J. (2018). John riccitiello sets out to identify the engine of growth for unity technologies (interview). In *Venture Beat. Interview with Dean Takahashi. Retrieved January*.

Schindler, G., & Dellaert, F. (2004). Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'04*.

Schonberger, J. L., & Frahm, J. M. (2016). Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'16* (pp. 4104–4113).

Singh, A., Patil, D., & Omkar, S. (2018) Eye in the sky: Real-time drone surveillance system (DSS) for violent individuals identification using scatternet hybrid deep learning network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, CVPRW'18*.

Sun, X., & Zheng, L. (2019). Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'19* (pp. 608–617).

Tardif, J. P. (2009). Non-iterative approach for fast and accurate vanishing point detection. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV'09* (pp. 1250–1257).

Vedaldi, A., & Zisserman, A. (2012). Self-similar sketch. In *Proceedings of the European Conference on Computer Vision, ECCV'12* (pp. 87–100).

Wang, T., Hu, X., Wang, Q., Heng, P. A., & Fu, C. W. (2020). Instance shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'20* (pp. 1880–1889).

Wildenauer, H., & Hanbury, A. (2012a). Robust camera self-calibration from monocular images of manhattan worlds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'12* (pp. 2831–2838).

Wildenauer, H., & Hanbury, A. (2012b). Robust camera selfcalibration from monocular images of Manhattan worlds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'12* (pp. 2831–2838).

Yang, W., Fang, B., & Tang, Y. Y. (2016). Fast and accurate vanishing point detection and its application in inverse perspective mapping of structured road. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 48*(5), 755–766.

Zhai, M., Workman, S., & Jacobs, N. (2016). Detecting vanishing points using global image context in a non-manhattanworld. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR'16* (pp. 5657–5665).

Zhang, S., Zhang, Q., Yang, Y., Wei, X., Wang, P., Jiao, B., & Zhang, Y. (2020). Person re-identification in aerial imagery. *IEEE Transactions on Multimedia, 23*, 281–291.

Zhao, J., Han, R., Gan, Y., Wan, L., Feng, W., & Wang, S. (2020). Human identification and interaction detection in cross-view multi-person videos with wearable cameras. In *Proceedings of the ACM International Conference on Multimedia, ACM MM'20* (pp. 2608–2616).

Zheng, K., Fan, X., Lin, Y., Guo, H., & Wang, S. (2017). Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras. In *Proceedings of the IEEE International Conference on Computer Vision, ICCV'17* (pp. 2858–2866).

Zhou, Y., Qi, H., Huang, J., & Ma, Y. (2019) NeurVPS: Neural vanishing point scanning via conic convolution. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS'19* (Vol. 32).