Augmenting Crowd-Sourced 3D Reconstructions using Semantic Detections

True Price[1] Johannes L. Schönberger[2] Zhen Wei[1] Marc Pollefeys[2,3] Jan-Michael Frahm[1]
[1]Department of Computer Science, UNC Chapel Hill [2]Department of Computer Science, ETH Zürich [3]Microsoft
{jtprice,zhenni,jmf}@cs.unc.edu {jsch,pomarc}@inf.ethz.ch

Abstract

Image-based 3D reconstruction for Internet photo collections has become a robust technology to produce impressive virtual representations of real-world scenes. However, several fundamental challenges remain for Structure-from-Motion (SfM) pipelines, namely: the placement and reconstruction of transient objects only observed in single views, estimating the absolute scale of the scene, and (suprisingly often) recovering ground surfaces in the scene. We propose a method to jointly address these remaining open problems of SfM. In particular, we focus on detecting people in individual images and accurately placing them into an existing 3D model. As part of this placement, our method also estimates the absolute scale of the scene from object semantics, which in this case constitutes the height distribution of the population. Further, we obtain a smooth approximation of the ground surface and recover the gravity vector of the scene directly from the individual person detections. We demonstrate the results of our approach on a number of unordered Internet photo collections, and we quantitatively evaluate the obtained absolute scene scales.

1. Introduction

Over the last decade, a significant effort has grown to actively map the 3D world around us into the virtual realm. In particular, virtual tourism applications [42] have become immensely popular in allowing users to experience and explore places they may otherwise not be able to visit.[1] These applications use 3D reconstruction approaches to recover models of actual places, typically starting from either a controlled capture scenario (*e.g.* aerial imagery of a single city) or, to visualize scenes from anywhere on Earth, publicly available photographs downloaded from the Internet. From this initial imagery, 3D models are obtained via Structure-from-Motion (SfM) [39, 43, 50, 49, 1, 12, 19, 40] and dense multi-view stereo (MVS) pipelines [41, 14, 13]. The ideal for virtual tourism is to present such reconstructions as *navigable environments* that a user can explore (*e.g.* in virtual reality) with a sense of "being" in the remote place.

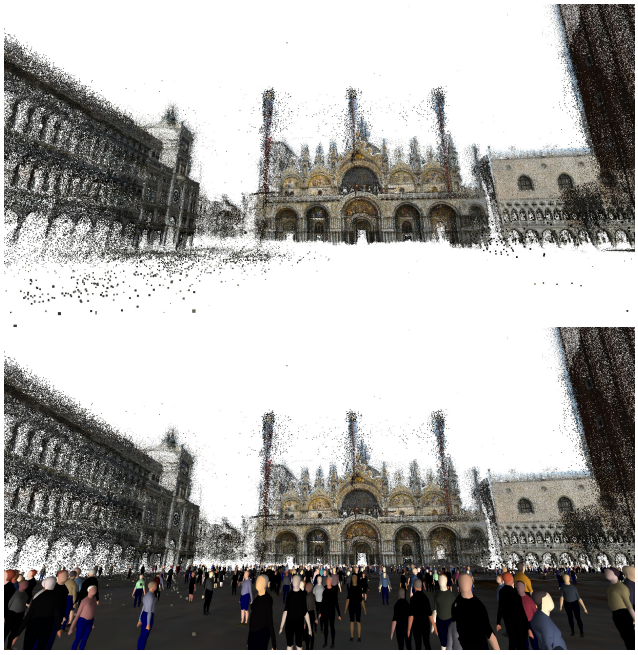

Figure 1. Result for our method for San Marco Square, Venice. Top: Dense, but incomplete, 3D reconstruction of static scene elements using multi-view stereo. Bottom: The same view, but with our textured ground surface added into the scene, and populated with a subset of pedestrians observed in the input images of the reconstruction. Our method jointly recovers the 3D position of the people, the absolute scale of the scene, the gravity direction of the scene, and the texture and geometry of the ground surface.

The perception of a virtual environment, however, is inherently tied to the *completeness* of the representation. Modern SfM+MVS pipelines have fundamental limitations in this respect. Ground reconstruction and transient object modeling are two bugbears of large-scale reconstruction from crowd-sourced still imagery. Ground is notoriously difficult to reconstruct from such image collections due to the relatively low number of matched ground points [26]. Transient objects such as people and moving cars are

[1] For example, Google Earth VR (https://vr.google.com/earth/) is a prominent virtual reality application that allows users to view cities from overhead in 3D.

likewise difficult to reconstruct without the availability of multiple synchronized views. In addition, automatically recovering the scene's absolute scale (*i.e*. reconstruction units per meter) is problematic unless a sufficient number of images have reliable world coordinate (*e.g*. GPS) information. Besides GPS' inherent unavailability for indoor scenes, the problem of missing GPS data is increasingly relevant given the recent trend of stripping out geo-location metadata from images shared online due to privacy concerns. Recovering these currently missing scene elements — transient objects, ground, and scale — is a highly desired goal in general 3D reconstruction, in addition to being a step forward for large-scale virtual tourism experiences.

In this paper, we seek to automatically augment SfM+MVS reconstructions with these fundamental scene elements. To achieve this, we leverage semantics on the transient objects in the scene. We specifically target large-scale 3D reconstructions obtained from Internet photo-collections, and we place people detected in the individual images into the 3D space; in principle, our method could be applied to other classes of transient objects, such as cars.

Our work provides a solution to an intriguing open problem in 3D reconstruction: How can moving objects be placed into a scene given only single observations from temporally disparate views? For large-scale SfM, the output 3D models are formed from many unordered images that effectively sample the behavior of people in the spatio-temporal domain. While the spatial domain is typically well-sampled because of the large number of images, the temporal sampling of the scene is mostly non-overlapping, *e.g*., on the order of a few (publicly shared) photos per hour even for highly photographed scenes. As a result, we can only assume that a single observation exists for any detection, which disallows the use of traditional triangulation approaches for 3D placement. To provide further context, we performed a cursory analysis of the date/time EXIF tags for several heavily photographed scenes in our experiments. In these datasets, we empirically observe that a publicly shared photo is taken approximately every 20 minutes.

To enable 3D placement, we leverage the fact that, given a sufficient number of observations and sufficiently visited areas, there will exist potentially many instances where multiple people in different images occupy the same location in the scene. Combined with viewing ray constraints and a known distribution of human height, this principle forms a strong cue for measuring the accuracy of 3D placement, which allows us to jointly recover scene scale, person placements, and ground surfaces. In this manner, we leverage object class semantics to "fill in" the parts of the reconstruction missed by traditional static methods. Fig. 1 shows an example of our reconstructed ground surfaces and 3D placement of people.

2. Related Work

There has been a strong interest in automatically obtaining 3D reconstructions from crowd-sourced images. The seminal work of Snavely *et al*. [42, 43] demonstrated the feasibility of reconstruction from Internet photos, and later systems robustified the reconstruction methods and tackled increasingly larger scenes and photo-collections. Today, state-of-the-art systems are able to provide highly detailed 3D models of thousands of sites around the world from one-hundred million user-uploaded images [19, 41]. However, the resulting models are only reconstructed up to an unknown scale factor and only represent the static parts of the scenes. Transient objects such as humans are inherently missing in such reconstructions.

A number of works have leveraged human detections for single-view camera calibration, particularly for surveillance cameras, and for crowd modeling in synchronized multi-view systems. Lv *et al*. [31, 32] and others [25, 23, 27, 34] extract head and foot positions for one or more walking humans in each frame of a video taken by a single stationary camera. Under the assumption that people stand upright and that the walking area is flat, these methods recover the vertical vanishing point and a horizon line for the scene, which can be further used to obtain camera intrinsics and the ground plane relative to the camera. If the height of one or more of the detected people is known, the absolute height of the camera above the ground can also be recovered. Notably, Liu *et al*. [29] used known human height distributions to automatically determine focal length and camera height. Other works [20, 46] explored increasing robustness by additionally incorporating vanishing points from the static scene. For general crowd modeling in multi-view synchronized systems [47], a large number of methods (*e.g*. [17, 10, 37, 11, 3]) exist to triangulate and track people in the camera space, potentially without explicit correspondences [30] or a knowledge of the system calibration [18]. Our work is set apart from these approaches in that we target the to-scale alignment of multiple camera spaces, which prevents the direct use of single-view methods, and that we cannot make use of a coherent scene configuration across views (or even repeated observations from a single view) in our temporally disjoint multi-view scenario.

Among other methods for reconstructing moving humans, trajectory triangulation for dynamic objects has been well-researched for images with dense temporal sampling [2, 38, 52, 22], but the topic has rarely been applied to unordered photo collections [53] and, to our knowledge, has not been applied in cases where hundreds or thousands of object class instances are observed. Garg *et al*. [16] explored detecting a single, manually specified individual among sets of Internet imagery, working under the assumption that the individual is positioned in approximately the same location across many images. Martin-Brualla

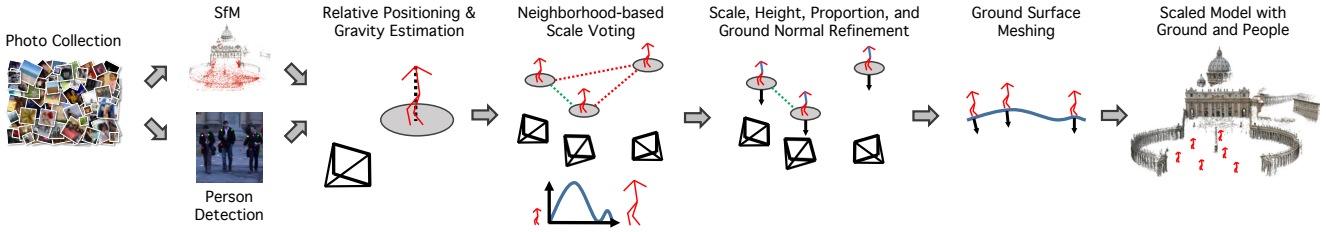

Figure 2. The pipeline of our proposed reconstruction system.

et al. [33] pieced together separate crowd-sourced 3D reconstructions by, in part, recovering the paths of photographers moving between them; this method does not recover the behavior of non-photographers, however. Zheng *et al*. [53] tackled the lack of temporal overlap by leveraging single-instance detections to localize object class trajectories. Their insight was that most object classes have structured motion paths in the scene, and recovering this path structure is complementary to recovering the object trajectories. The problem is formulated as a generalized minimum spanning tree (GMST), followed by a continuous optimization to refine the trajectory. However, the approach does not generalize to unstructured or weakly structured object class motions, as is often encountered in open scenes such as plazas or tourist sites. Additionally, their method carries high computational cost due to solving the NP-hard problem of computing the GMST [35]. In contrast, our method dispenses with the requirement of structured object class motion along a path. Moreover, it is significantly more efficient in terms of computation, allowing us to scale to large-scale photo-collections with thousands of images per site. We achieve this by incorporating semantic information and coarse object class triangulation.

Similar to our work, Bulbul and Dahyot [5] introduced a method for obtaining representations of transient objects in map representations such as OpenStreetMap (OSM) [36]. In contrast to our scenario, OSM provides both a to-scale, geo-localized environment model and a coarse ground surface representation. The authors used social media photos with geo-localization metadata to place human avatars into the map. To obtain the camera position of a social media image, they registered the image to nearby Google StreetView images[2] based on its known geo-location. People in the images were placed onto the map's ground surface at a distance from the camera estimated by the size of their face in the image. In this paper, we seek to not only place people into 3D models of places on Earth, but to actually complete reconstructions in terms of scale, transient objects, and ground, all without any external references.

[2] https://www.google.com/streetview/

3. Methods

In this section, we present our novel approach for placing people, estimating scale, and recovering ground surface in a 3D scene. An overview of our pipeline is shown in Fig. 2. Starting from an initial set of photos of a scene, we first employ Structure-from-Motion (SfM) [39] to obtain camera parameters and sparse structure. We then detect 2D torso points for people in the images [48, 7] and from these estimate the distances and rotations of individuals relative to each camera, as well as a global scene gravity vector (Section 3.1). We next test out a range of possible scene scales for the reconstruction and rank them using approximate semantic triangulation (Section 3.2). We then refine the scale and the 3D placement of the people using known human height statistics and encouraging a locally planar ground surface (Section 3.3). In the last stage, we recover the ground surface using Poisson surface reconstruction [24] (Section 3.4). For visualization, we place human avatars into the 3D space with clothing colors sampled from the input images; we also texture the ground using image data and semantic pixel labelings [51] (Section 3.5).

3.1. Person Detection and Gravity Estimation

The input to our algorithm consists of a set of photos of a scene, plus a sparse representation of the scene obtained from these images via SfM [39]. Our first step is to detect people in the images and obtain an initial estimate of each person's *absolute position* – that is, the real-world coordinates (in meters) of the person in the reference frame of the camera when the image was taken. These initial positions will subsequently be used for a coarse scene scale estimation. The general approach we take here is to detect torso points in each image and, for each detection, fit a planar torso model to the detected points. We assume that detected torsos are aligned with the (initially unknown) gravity vector for the scene, which is a generally valid assumption given that most people stand upright [31, 25, 34]. We jointly optimize 1) the global gravity vector, 2) the absolute position of each person's neck point, and 3) the 1-DoF heading (rotation around the gravity vector) of the person. This optimization is done by minimizing the reprojection error of the posed torso models back into their original images.

Torso Detection: For detection, we use Convolutional

Pose Machines (CPM) [48, 7], a state-of-the-art joint detector specifically designed for real-time, multi-person pose estimation. We define the 2D joints on the torso by taking the CPM detections for the neck, shoulders, and hips. We only consider joint detections having at least 30% confidence, and we rule out individuals lacking confident detections in the neck and at least one of the hips.

Torso Model Fitting: As a coarse initialization that will later be refined, we fit a fixed-size planar torso model to each detection. This model is centered at the neck point with a width of 30cm and a height of 52cm (Fig. 3). Our convention is that gravity points in the positive y direction, so the model is defined in the xy plane.

We transform the torso model to match the detected 2D joints for person i. Because we have obtained an initial SfM reconstruction of the scene, we know the pose $[R_i \mid \mathbf{t_i}]$ and the intrinsics of the observing camera. The camera location in the reconstruction space does not matter at this stage, but it is necessary to know the orientation of the camera relative to the gravity direction of the scene.

We apply the model-to-camera transformation in four steps. First, we rotate the model around the y axis by angle θ_i; denote the associated rotation as $R(\theta_i)$. This represents the direction the person is facing in the reconstruction space. Second, we align the model to the scene gravity vector $\mathbf{g} \in \mathbb{R}^3$, with $||\mathbf{g}|| = 1$, by calculating the rotation of the model gravity vector $[0\ 1\ 0]^T$ into $\mathbf{g}$. This rotation can be formulated as the unit quaternion $\mathbf{q_g} = (\hat{v}_2, \hat{v}_3, 0, -\hat{v}_1)$, where $\hat{v} = \frac{v}{||v||}$ with $v = \mathbf{g} + [0\ 1\ 0]^T$; more generally, we denote this model-to-world gravity alignment as $R(\mathbf{g})$. Third, we place this result in the coordinate frame of the observing camera by applying the extrinsic rotation matrix R_i. Finally, we translate the model relative to the camera based on the 3D position of the neck point $N_i = z_i[x_i\ y_i\ 1]^T$, where (x_i, y_i) is the 2D coordinate of the neck point in normalized camera coordinates, and z_i is the depth (in meters) of the person relative to the camera. Note, we do not require (x_i, y_i) to exactly lie at the neck point detected by CPM.

For 3D joint J_m in the original torso model, we thus obtain a rotated, gravity-aligned, camera-aligned 3D joint:

$$J_{i,m} = R_i R(\mathbf{g}) R(\theta_i) J_m + N_i. \tag{1}$$

Optimization: We jointly optimize $\mathbf{g}$ and all individuals' poses $\mathbf{\Theta} = \{(\theta_i, x_i, y_i, z_i)\}$ by minimizing the reprojection errors of the torso model into the original images:

$$\min_{\mathbf{g},\mathbf{\Theta}} \sum_i \phi \left(\sum_m \rho_{i,m}^2 ||\pi_i(J_{i,m}) - \mathbf{j_{i,m}}||^2 \right), \tag{2}$$

where $\mathbf{j_{i,m}}$ is the 2D pixel location of detected joint m, $\pi_i(\cdot)$ is the projection function for camera i that converts 3D points relative to the camera into 2D pixel projections according to the camera intrinsics estimated in SfM, and

Figure 3. To accurately localize 2D ground points for detected people, we first fit a planar torso model in 3D (left) to detected 2D neck, shoulder, and hip joints (middle-left). Right: Coordinate axes for the planar model.

$\rho_{i,m}$ is the joint detection confidence obtained from CPM. $\phi(\cdot)$ is a robust function that mitigates the effect of strong outlier detections; in our implementation, we employ the Huber loss function with a threshold of 4 pixels [21].

The gravity vector is initialized to the geometric median of the individual camera down vectors. In order to obtain good initialization for depth, we perform a preliminary optimization of depths $\{z_i\}$ and gravity only, followed by a further optimization of all parameters. The depth/gravity optimization works as follows: Neck locations $\{(x_i, y_i)\}$ are fixed to the initially detected 2D locations, and depths are initialized to 1 meter. The rotation parameters $\{\theta_i\}$ are ignored; instead, we sample a set of discrete rotations $\{\bar{\theta}_k\}$ at intervals of 10°. For each detection, the optimal rotation is taken as the angle in this set that minimizes the reprojection error. A modified version of Eq. (2) is thus optimized:

$$\min_{\mathbf{g},\{z_i\}} \sum_i \phi \left(\min_{\bar{\theta}_k} \sum_m \rho_{i,m}^2 ||\pi_i(J_{i,m}(\bar{\theta}_k)) - \mathbf{j_{i,m}}||^2 \right), \tag{3}$$

where $J_{i,m}(\bar{\theta}_k) = R_i R(\mathbf{g}) R(\bar{\theta}_k) J_m + N_i$.

After this first optimization, $\{\theta_i\}$ values are initialized based on the value of $\bar{\theta}_k$ that minimizes the reprojection error for each person. The full set of parameters $(\mathbf{g}, \mathbf{\Theta})$ is then optimized using Eq. (2). Finally, we re-orient the 3D reconstruction such that the estimated gravity vector is aligned with the positive y axis.

3.2. Voting-based Scale Estimation

At this point, we have obtained an initial absolute depth estimate for each person relative to the camera that observes them. Next, we estimate an initial placement of the detections into the reconstruction space, while at the same time obtaining an initial absolute scale estimate for the scene. If the scene scale s (*e.g.* the length of 1 meter in the reconstruction space) were known, we could calculate the 3D neck point of person i in the reconstruction space as

$$P_i(s) = s R_i^T N_i + C_i, \tag{4}$$

where $N_i \in \mathbb{R}^3$ is the estimated 3D position of the neck point relative to the observing camera, $R_i \in \mathbb{R}^{3\times3}$ is the scene-to-camera rotation matrix, and $C_i \in \mathbb{R}^3$ is the 3D position of the camera in the reconstruction space.

In principle, s could be determined from a known absolute distance between two points in the reconstruction space, *e.g.*, the width of a building or the distance between two cameras. Lacking known distances, we propose to instead leverage *approximate semantic triangulation*. The idea here is that, given enough input images, and especially in well-traveled areas, there is a high probability that at least two individuals in different images will be observed in nearby locations, and at similar heights above the ground. Our method samples a range of scale hypotheses for the 3D reconstruction and scores each based on the observed person correspondences.

Pairwise Approximate Triangulation: More explicitly, consider the 3D neck placements $P_i(s)$ and $P_j(s)$ (Eq. (4)) for two individuals at some scene scale s. Recall that, by convention, the y axis defines the vertical span of the scene, and the xz plane defines the horizontal space. We denote two individuals as standing "nearby" if they are within some fixed absolute distance τ_{xz} in the horizontal space. In addition, we say that the individuals are standing at similar heights if their neck points are within some fixed absolute distance τ_y in the vertical space. Taking $\Delta P_{ij}(s) = P_i(s) - P_j(s)$, let $M_{ij}(s)$ denote the binary indicator function that determines whether persons i and j are approximately triangulated at scale s:

$$M_{ij}(s) = \left(||\Delta P_{ij}^{xz}(s)|| < s\tau_{xz}\right) \wedge \left(|\Delta P_{ij}^{y}(s)| < s\tau_y\right), \quad (5)$$

where $||\Delta P_{ij}^{xz}(s)||$ and $|\Delta P_{ij}^{y}(s)|$ denotes the horizontal and vertical distances between the neck points, respectively.

We compute $M_{ij}(s)$ for all pairs of detected people in separate images. We also only consider individuals satisfying visibility constraints ($V_i(s)$, explained below). An individual is successfully triangulated at scale s if any pairwise approximate triangulation was successful:

$$M_i(s) = V_i(s) \wedge \left(\bigvee_j \left(\mathcal{I}_i \neq \mathcal{I}_j\right) \wedge V_j(s) \wedge M_{ij}(s) \right), \quad (6)$$

where $\mathcal{I}_i$ denotes the image in which person i was detected.

Visibility Constraint: An important constraint in our scale estimation is that the line segment from C_i to $P_i(s)$ should not intersect with structures such as walls. This constraint may be violated if s is too large, which pushes $P_i(s)$ further from the observing camera. Accordingly, $V_i(s)$ is an indicator function denoting whether the detection of person i is possible at scale s given the free space of the static parts of the scene. In practice, we compute $V_i(s)$ by voxelizing the SfM 3D point cloud with a fixed voxel size of one meter

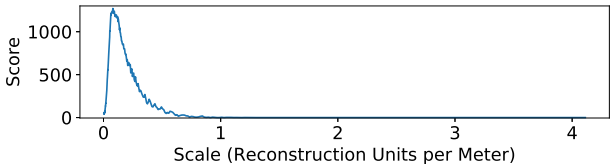

Figure 4. Scale scoring curve for our model of the Pantheon. The peak is chosen as our initial scale estimate.

(s units in the reconstruction space). We then perform ray-tracing from C_i along ray $R_i^T N_i$ and compute the first point of intersection with a filled voxel. We denote the distance from C_i to this voxel as $v_i(s)$. $V_i(s)$ is then defined as

$$V_i(s) = s||N_i|| < v_i(s). \quad (7)$$

Scale Scoring: We score a hypothesized scale s by taking a weighted aggregate of all $M_i(s)$:

$$S(s) = \sum_i w_i M_i(s). \quad (8)$$

Setting $w_i = 1$ is equivalent to counting the successfully triangulated individuals at scale s. We experimentally found better performance by weighting individuals by the number of detections in their associated image, *i.e.*, $w_i = 1/N_{\mathcal{I}_i}$, where $N_{\mathcal{I}_i}$ is the total number of detections in image $\mathcal{I}_i$. This weighting mitigates the ambiguity of person placement in crowded areas, where incorrect scales can still yield valid triangulations due to the overall person density.

Finally, we obtain an initial voting-based estimate of the scene scale by sampling a range of possible scales and selecting the scale hypothesis with the highest score $S(s)$. In our experiments, we generate this range by assuming that the vertical span of the SfM point cloud is between 1 and 1000 meters. We start at the smallest possible scale and test all scales in the range, stepping at 2% increments in s. At this stage, we only consider individuals having all five torso joints detected with at least 30% confidence. We use absolute horizontal and vertical thresholds of $\tau_{xz} = 1.5$m and $\tau_y = 0.1$m. An example scoring curve is shown in Fig. 4.

3.3. Scale Refinement, Height Estimation, and Ground Surface Estimation

Having obtained an initial scale estimate s, we next jointly refine this scale, estimate a height h_i in meters for each detected individual, and estimate a ground surface unit normal $\mathbf{n_i} \in S^2$ for the ground point at which each individual stands. As part of this optimization, we also estimate a torso height $t_i = \beta_i h_i$ for each person, where β_i is the individual's torso-to-height proportion. In the following, we first formulate how to obtain a person's 3D position in the reconstruction space given s, h_i, and β_i. We then introduce the three terms of our joint optimization function and finally address the overall formulation.

Position as a Function of Height and Proportion: While Eq. (4) is convenient for an initial neck point placement, it relies on a fixed torso size. We generalize this formula by allowing the torso height t_i to vary as a fraction β_i of the person's height h_i. The end result is that an increase or decrease in torso size accordingly affects the distance of the neck point N_i in Eq. (4) to the camera.

Let $\mathbf{r_i} = N_i/||N_i||$ denote the ray from the origin through the neck point of the fitted torso model in the reference frame of the camera. Moreover, let $\mathbf{h_i}$ be the ray for the hip midpoint of the model. For every 3D point falling on $\mathbf{r_i}$, there is an associated point on $\mathbf{h_i}$ that falls directly below it along the gravity direction (y axis). Again assuming that the torso aligns with the gravity vector, we can find such a neck/hip point pair for any torso height t_i. By similar triangles, we can determine a new neck point $N_i(t_i) = \varrho_i t_i \mathbf{r_i}$ for any torso height, where ϱ_i is the the ratio between neck-point-to-camera distance and torso height.

In practice, we explicitly encode an understanding of human proportions by expressing torso height as a percentage of total height, *i.e.*, $t_i = \beta_i h_i$. We can thus update Eq. (4) to express a person's 3D neck point in the reconstruction space (at scale s) as a function of height and proportion:

$$P_i(s, h_i, \beta_i) = sR_i^T N_i(t_i) + C_i = \varrho_i \beta_i h_i R_i^T \mathbf{r_i} + C_i, \quad (9)$$

We also include photographers into the optimization in Eq. (14). However, since we do not observe torsos for photographers, we must treat them slightly differently. Specifically, we assume that the camera center is $h_c/8$ meters above the neck point for photographer c. Accordingly, $\mathbf{r_c} = [0\ 1\ 0]$, and we fix $\varrho_c = 1$ and $\beta_c = 1/8$.

Ground Point Position: The ground point $G_i(s, h_i)$ lies vertically below the neck point $P_i(s, h_i, \beta_i)$. With the neck height being a fraction η of the total height of the person, the ground point in reconstruction space is given as

$$G_i(s, h_i, \beta_i) = P_i(s, h_i, \beta_i) + [0\ s\eta h_i\ 0]^T. \quad (10)$$

We use a fixed value of $\eta = 5/6$, reasoning that the top of the sternum (our assumed neck point) is slightly less than two head lengths from the top of a person, and that human head length is approximately one-eighth of total height [4].

Optimization Overview: As previously mentioned, we optimize scale s along with the set $\{(h_i, \beta_i, \mathbf{n_i})\}$ of per-person heights, proportions, and ground normals. Our objective function has three terms: 1) a prior on height, 2) a local ground planarity term for pairs of nearby people, and 3) a visibility constraint.

Height Distribution Prior: We propose to leverage the known distribution of human heights as a prior on the estimated height h_i for each person. Here, we employ a Gaussian mixture model (GMM) for this distribution; in principle, any GMM or otherwise appropriate probability distribution could be used. The GMM probability function is given as the sum of probabilities for K separate Gaussians:

$$p(h_i) = \sum_{k=1}^{K} \frac{\alpha_k}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{(h_i - \mu_k)^2}{2\sigma_k^2}\right), \quad (11)$$

Here, we use a general two-component GMM for male and female adult heights, respectively: $\{(\alpha_k, \mu_k, \sigma_k)\} = [(0.504, 1.768, 0.068), (0.496, 1.646, 0.060)]$, which we aggregated from several sources [15, 44, 45]. In principle, more detailed models could be used, such as a model that captures factors of age or ethnicity.

Local Planarity Prior: Our second objective term encourages the ground surface between nearby people to be relatively smooth (but not necessarily horizontal). We enforce this by endowing each individual with a ground normal $\mathbf{n_i}$ that defines a planar ground patch around the point at which they stand. We penalize nearby ground points that are far from this flat surface. For two individuals i and j, the point-to-plane distance in meters between $G_j(s, h_j, \beta_j)$ and the patch for person i is given as

$$d_{ij} = \frac{1}{s} \left| \left(G_j(s, h_j, \beta_j) - G_i(s, h_i, \beta_i)\right)^T \mathbf{n_i} \right|. \quad (12)$$

Visibility Constraint: We again seek to penalize scales and heights that push neck points into or beyond static parts of the scene. To do this, we repeat our earlier voxelization at the initial scale s_0 and compute $v_i(s_0)$. These maximum distances are then fixed in our optimization. Our penalty term for this is close to zero for neck-to-camera distances much less than v_i and close to one for values much greater:

$$\nu_i(s, h_i, \beta_i) = \frac{1}{\pi \tan^{-1}(2)} \tan^{-1}\left(\frac{2}{\tau_o}\left(||N_i(t_i)|| - \frac{v_i}{s}\right)\right), \quad (13)$$

where $||N_i(t_i)|| = \varrho_i \beta_i h_i$ is the neck-to-camera distance in meters, and τ_o is a value in meters such that an "overshooting" of $3\tau_o$ meters results in a penalty of approximately 0.95. In our experiments, we use τ_o = 0.2m.

Optimization: We combine Eqs. (11-13) into a single objective function to be minimized:

$$E(s, \{(h_i, \beta_i, \mathbf{n_i})\}) = -\frac{1}{D}\sum_{i=1}^{D} \log p_i(h_i) + \frac{1}{4|\mathcal{N}|\lambda^2} \sum_{(i,j)\in\mathcal{N}} \left(d_{ij}^2 + d_{ji}^2\right) + \frac{1}{D}\sum_{i=1}^{D} \nu_i(s, h_i, \beta_i), \quad (14)$$

where D is the total number of detected people, $\mathcal{N}$ is a set of person neighbors to which the local planarity prior is applied bidirectionally, and λ is a penalty term for the planarity penalty. The first term was derived by taking the negative log-likelihood of the height probability. In our experiments, we set $\lambda = 0.02$, which roughly reflects an expected

ground plane noise of 2cm. We define the neighborhood structure $\mathcal{N}$ based on our initial person placements at s_0. We classify nearby initial placements as those having neck points within 3m of each other in the horizontal space and 0.242m in the vertical space. Under our height model, the vertical threshold is the point at which 95% of randomly chosen height pairs are expected to fall within that value.

We constrain $\beta_i \in [0.25, 0.45]$, which reasonably captures the range of human torso proportions [4], and we initialize these values to 0.3 for optimization. We initialize individual heights randomly by sampling from our height distribution model. Normals are parameterized by spherical coordinates, which we initialize with small random perturbation. At this stage, we also include person detections having at least four detected joints.

3.4. Ground Surface Reconstruction

Using the optimized 3D ground points and ground point normals, we fit a ground surface using the Poisson surface reconstruction (PSR) implementation of Kazhdan and Hoppe [24]. PSR produces a high-quality mesh with adaptive resolution from an input set of oriented points, which in our case is defined by $\{ (G_i(s, h_i, \beta_i), \mathbf{n}_i) \}$. Prior to running PSR, we filter the input point cloud by removing individuals who are more than 40m from their observing camera or who fail the visibility constraint at the optimized scale s, and we also remove small, far-off groups of photographers.

3.5. Visualization

To demonstrate the potential of our method for scene completion, we texture our recovered ground surface and place a subset of all detected people into the reconstruction space. Our person visualization consists of a low-poly model for each detection with the shirt and pants colored by sampling the original image. Each person model is scaled to match our estimated height for the detection. Due to the large number of detections in many of our scenes, we choose a subset of individuals by treating the selection as a set cover problem and taking a greedy approach. Specifically, for each photographer c, we denote $\mathcal{O}_c$ as the set of people observed in the image taken by that cameraperson, and $\mathcal{V}_c \supseteq \mathcal{O}_c$ as the set of all individuals (including photographers) placed within the viewing frustum of the photographer's camera, up to some maximum depth. We select a photographer and mark all individuals in $\mathcal{V}_c$ as "visited." At the same time, we place in the reconstruction all individuals in $\mathcal{O}_c$ who were not previously marked as visited; if any such person exists, the photographer is also placed into the scene. We randomly and iteratively select photographers in this fashion until all people are marked as visited.

Additionally, we texture the ground surface obtained from PSR, which has accurate geometry but lacks color. For each vertex on the ground surface mesh, we project the vertex into each individual image in our 3D reconstruction and, if the projection lies within the image boundaries, sample the color value at the pixel in which it falls. We aggregate the sampled colors over all images and take the median color for each vertex. To avoid sampling non-ground pixels (caused by, *e.g.*, occluding scene geometry or pedestrians), we leverage recent advances in dense pixel-wise semantic labeling. For each image, we apply the convolutional neural network of [51], trained on the Cityscapes dataset [8], to obtain a most-probable class labeling for each pixel. When aggregating color values, we ignore sampled pixels that do not receive class labels of ground, sidewalk, or terrain.

4. Results

We have tested our method on several large-scale image photo-collections [28, 6, 49], as well as the well-known Cornell Arts Quad dataset [9]. Evaluation in the context of unordered Internet photo-collections is a challenging task due to the lack of available ground truth. Hence, we manually establish ground truth for the scale estimation of our method by obtaining known distances of structures in the scene, which are then compared to the same distance in our model. The distance evaluation results are shown in Table 1. It can be seen that our proposed scale estimation via a height distribution prior reliably determines the scene scale. Effectively, our method uses object semantics to overcome the inherent scale ambiguity of SfM reconstructions, which has long been a goal of computer vision.

We additionally evaluated our gravity vector estimation for the scenes from [49] and found an average error of 1.078° when compared to the implementation of automatic gravity vector estimation from scene vanishing points in [39]. Given the lack of obtainable ground-truth for the 3D placement of people from a particular photo, we qualitatively evaluate the placement using randomly selected

Scene	**Error**	n_p	n_c
Cornell Quad	-4.0%	550	4773
Dubrovnik	-0.15%	5066	2714
Pantheon	+4.3%	8656	3310
Campitelli	+1.9%	15836	16834
San Marco	-0.3%	15712	4916
Alamo	+0.3%	1940	699
NYC Lib.	-1.4%	466	480
Piccadilly	-6.1%	7908	2453

Table 1. Quantitative results on our method for scale and placement. "% Error" gives the amount that we over/under-estimated the distance of one unit in the reconstruction. n_p and n_c show the number of placed detected people and photographers, respectively, recovered by our method.

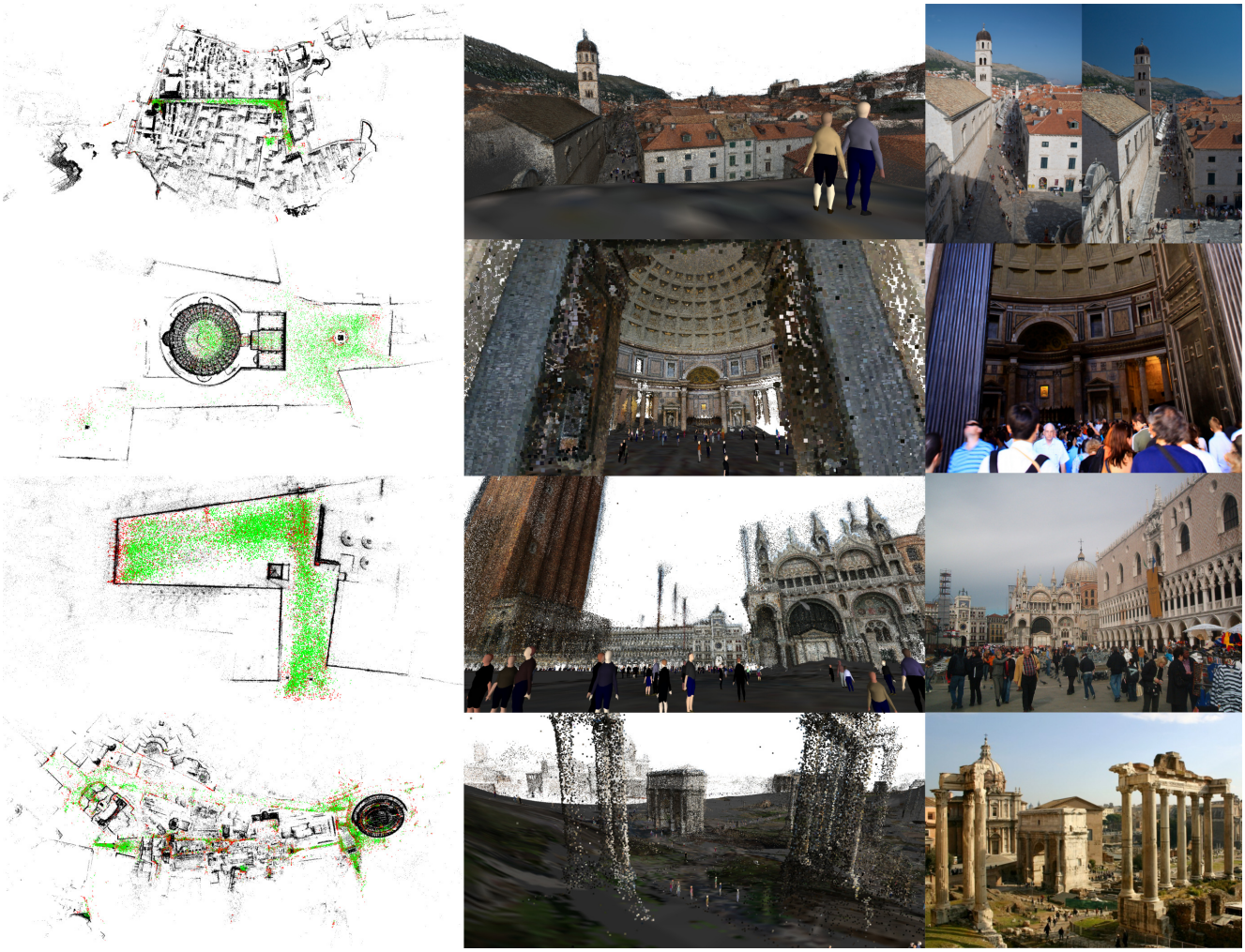

Figure 5. Overhead views (left) and sample renderings with ground and person avatars (middle) for our method. Examples of real photos are shown on the right. The green dots in the overhead views show person placements, with cameras as red dots and detected people as green dots. Black dots show static structure. From top: Dubrovnik, Croatia; the Pantheon; San Marco Plaza, Venice; and the area around the Colosseum and Roman Forum in Rome.

photos from the photo collection. Sample evaluations are shown in Fig. 5 on four large-scale datasets: Dubrovnik, the Pantheon, San Marco Plaza, and the Campitelli in Rome. Additional results, including ablative analyses of the steps of our method, are available in the supplementary material. In summary, we demonstrated the reliable and accurate behavior of our proposed method for bringing scenes to life. We provide evaluations that prove our solutions to the challenging open problems of obtaining accurate scaling for a reconstructed scene, correctly placing transient objects, and estimating ground surface from unordered Internet photo-collections, which has not been previously solved at scale.

5. Conclusion

We have introduced a new approach for adding living, transient elements to large-scale static 3D reconstructions. Specifically, our method leverages recent advances in image-based person detection, along with population height distribution priors, to jointly place detected people into the scene, estimate the absolute scale of the reconstruction, recover the gravity vector of the scene, and recover the underlying ground surface. We have tested our method on a large collection of real-world datasets and demonstrate quantitative and qualitative results that verify the significant advances of our approach in modeling hard-to-capture scene elements. A key insight of our work is that knowledge of object class properties, such as height distribution in humans, can provide adequate constraints on 3D placement even when exact correspondence is impossible. In the future, we look to extend our work to areas such as crowd simulation and integrated processing with online videos.

Acknowledgements This research was partially supported by NSF grant No. CNS-1405847.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 1

[2] S. Avidan and A. Shashua. Trajectory triangulation: 3d reconstruction of moving points from a monocular image sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):348–357, 2000. 2

[3] J. Black, T. Ellis, and P. Rosin. Multi view image surveillance and tracking. In *Workshop on Motion and Video Computing*, pages 169–174. IEEE, 2002. 2

[4] B. Bogin and M. I. Varela-Silva. Leg length, body proportion, and health: a review with a note on beauty. *International journal of environmental research and public health*, 7(3):1047–1075, 2010. 6, 7

[5] A. Bulbul and R. Dahyot. Populating virtual cities using social media. *Computer Animation and Virtual Worlds*, 2016. 3

[6] S. Cao and N. Snavely. Learning to match images in large-scale collections. In *European Conference on Computer Vision (ECCV), Workshops and Demonstrations*, pages 259–270. Springer, 2012. 7

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 4

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[9] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3001–3008. IEEE, 2011. 7

[10] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, 2008. 2

[11] D. Focken and R. Stiefelhagen. Towards vision-based 3-d people tracking in a smart room. In *International Conference on Multimodal Interfaces*, pages 400–405. IEEE, 2002. 2

[12] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, et al. Building rome on a cloudless day. In *European Conference on Computer Vision (ECCV)*, pages 368–381. Springer, 2010. 1

[13] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1434–1441. IEEE, 2010. 1

[14] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010. 1

[15] J. Garcia and C. Quintana-Domeque. The evolution of adult height in europe: a brief note. *Economics & Human Biology*, 5(2):340–349, 2007. 6

[16] R. Garg, S. M. Seitz, D. Ramanan, and N. Snavely. Where's waldo: Matching people in images of crowds. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1793–1800. IEEE, 2011. 2

[17] W. Ge and R. T. Collins. Crowd detection with a multiview sampler. In *European Conference on Computer Vision (ECCV)*, pages 324–337. Springer, 2010. 2

[18] J. Guan, F. Deboeverie, M. Slembrouck, D. Van Haerenborgh, D. Van Cauwelaert, P. Veelaert, and W. Philips. Extrinsic calibration of camera networks based on pedestrians. *Sensors*, 16(5):654, 2016. 2

[19] J. Heinly, J. L. Schönberger, E. Dunn, and J.-M. Frahm. Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset). In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[20] M. Hödlmoser, B. Micusik, and M. Kampel. Camera autocalibration using pedestrians and zebra-crossings. In *International Conference on Computer Vision (ICCV) Workshops*, pages 1697–1704. IEEE, 2011. 2

[21] P. J. Huber. *Robust Statistics*. John Wiley and Sons, 1981. 4

[22] D. Ji, E. Dunn, and J.-M. Frahm. 3d reconstruction of dynamic textures in crowd sourced data. In *European Conference on Computer Vision (ECCV)*, pages 143–158. Springer, 2014. 2

[23] I. Junejo and H. Foroosh. Robust auto-calibration from pedestrians. In *International Conference on Video and Signal Based Surveillance (AVSS)*, pages 92–92. IEEE, 2006. 2

[24] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)*, 32(3):29, 2013. 3, 7

[25] N. Krahnstoever and P. R. Mendonca. Bayesian autocalibration for surveillance. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 1858–1865. IEEE, 2005. 2, 3

[26] A. Kuhn, T. Price, J.-M. Frahm, and H. Mayer. Down to earth: Using semantics for robust hypothesis selection for the five-point algorithm. In *German Conference on Pattern Recognition (GCPR)*, pages 389–400. Springer, 2017. 1

[27] W. Kusakunniran, H. Li, and J. Zhang. A direct method to self-calibrate a surveillance camera by observing a walking pedestrian. In *Digital Image Computing: Techniques and Applications*, pages 250–255. IEEE, 2009. 2

[28] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *European Conference on Computer Vision (ECCV)*, pages 791–804. Springer, 2010. 7

[29] J. Liu, R. T. Collins, and Y. Liu. Surveillance camera autocalibration based on pedestrian height distributions. In *British Machine Vision Conference (BMVC)*, 2011. 2

[30] J. Liu, R. T. Collins, and Y. Liu. Robust autocalibration for a surveillance camera network. In *Workshop on Applications of Computer Vision (WACV)*, pages 433–440. IEEE, 2013. 2

[31] F. Lv, T. Zhao, and R. Nevatia. Self-calibration of a camera from video of a walking human. In *International Conference on Pattern Recognition (ICPR)*, volume 1, pages 562–567. IEEE, 2002. 2, 3

[32] F. Lv, T. Zhao, and R. Nevatia. Camera calibration from video of a walking human. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1513–1518, 2006. 2

[33] R. Martin-Brualla, Y. He, B. C. Russell, and S. M. Seitz. The 3d jigsaw puzzle: Mapping large indoor spaces. In *European Conference on Computer Vision (ECCV)*, pages 1–16. Springer, 2014. 3

[34] B. Micusik and T. Pajdla. Simultaneous surveillance camera calibration and foot-head homology estimation from human detections. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1562–1569. IEEE, 2010. 2, 3

[35] Y.-S. Myung, C.-H. Lee, and D.-W. Tcha. On the generalized minimum spanning tree problem. *Networks*, 26(4):231–241, 1995. 3

[36] OpenStreetMap contributors. Planet dump retrieved from http://planet.osm.org. http://www.openstreetmap.org, 2017. 3

[37] K. Otsuka and N. Mukawa. Multiview occlusion analysis for tracking densely populated objects based on 2-d visual angles. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–I. IEEE, 2004. 2

[38] H. S. Park, T. Shiratori, I. Matthews, and Y. Sheikh. 3d reconstruction of a moving point from a series of 2d projections. In *European Conference on Computer Vision (ECCV)*, pages 158–171. Springer, 2010. 2

[39] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3, 7

[40] J. L. Schönberger, F. Radenović, O. Chum, and J.-M. Frahm. From Single Image Query to Detailed 3D Reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[41] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[42] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM transactions on graphics (TOG)*, volume 25, pages 835–846. ACM, 2006. 1, 2

[43] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008. 1, 2

[44] S. Subramanian, E. Özaltin, and J. E. Finlay. Height of nations: a socioeconomic analysis of cohort differences and patterns among women in 54 low-to middle-income countries. *PLoS One*, 6(4):e18962, 2011. 6

[45] The World Bank Group. Population, female (% of total). https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS. Accessed: 2017-11-15. 6

[46] T. Trocoli and L. Oliveira. Using the scene to calibrate the camera. In *SIBGRAPI Conference on Graphics, Patterns and Images*, pages 455–461. IEEE, 2016. 2

[47] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1):3–19, 2013. 2

[48] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4

[49] K. Wilson and N. Snavely. Robust global translations with 1dsfm. In *European Conference on Computer Vision (ECCV)*, 2014. 1, 7

[50] C. Wu et al. Visualsfm: A visual structure from motion system. 2011. 1

[51] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 3, 7

[52] E. Zheng, D. Ji, E. Dunn, and J.-M. Frahm. Sparse dynamic 3d reconstruction from unsynchronized videos. In *International Conference on Computer Vision (ICCV)*, pages 4435–4443, 2015. 2

[53] E. Zheng, K. Wang, E. Dunn, and J.-M. Frahm. Joint object class sequencing and trajectory triangulation (jost). In *European Conference on Computer Vision (ECCV)*, 2014. 2, 3