Virtual Correspondence: Humans as a Cue for Extreme-View Geometry

Wei-Chiu Ma[1,4] Anqi Joyce Yang[2,4,5] Shenlong Wang[3] Raquel Urtasun[2,4,5] Antonio Torralba[1]
[1]Massachusetts Institute of Technology [2]University of Toronto
[3]University of Illinois Urbana-Champaign [4] Waabi [5] Vector Institute

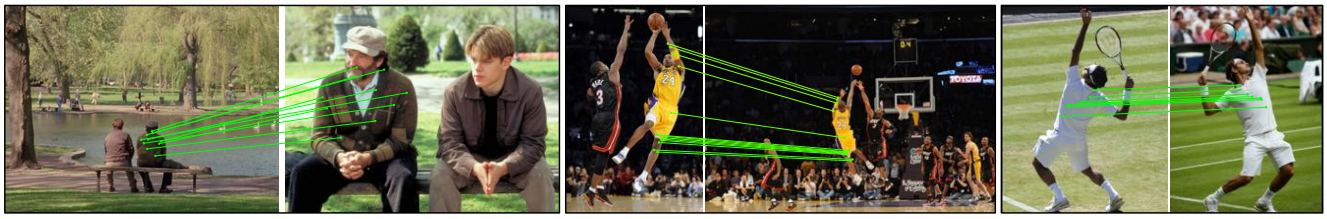


Figure 1. **Can you tell the relationships between these matched pixels?** The head pixel and the face pixel in the leftmost images have completely different semantics and appearances, yet we can still associate them for 3D reasoning. Why, and how? In this paper, we present a novel concept to establish geometric relationships between pixels even if they are not semantically or visually similar. See Fig. 11 for cars.

Abstract

Recovering the spatial layout of the cameras and the geometry of the scene from extreme-view images is a long-standing challenge in computer vision. Prevailing 3D reconstruction algorithms often adopt the image matching paradigm and presume that a portion of the scene is co-visible across images, yielding poor performance when there is little overlap among inputs. In contrast, humans can associate visible parts in one image to the corresponding invisible components in another image via prior knowledge of the shapes. Inspired by this fact, we present a novel concept called virtual correspondences (VCs). VCs are a pair of pixels from two images whose camera rays intersect in 3D. Similar to classic correspondences, VCs conform with epipolar geometry; unlike classic correspondences, VCs do not need to be co-visible across views. Therefore VCs can be established and exploited even if images do not overlap. We introduce a method to find virtual correspondences based on humans in the scene. We showcase how VCs can be seamlessly integrated with classic bundle adjustment to recover camera poses across extreme views. Experiments show that our method significantly outperforms state-of-the-art camera pose estimation methods in challenging scenarios and is comparable in the traditional densely captured setup. Our approach also unleashes the potential of multiple downstream tasks such as scene reconstruction from multi-view stereo and novel view synthesis in extreme-view scenarios[1].

[1]Project page: https://people.csail.mit.edu/weichium/virtual-correspondence/

1. Introduction

Epipolar geometry and correspondence estimation are two keystones of mainstream 3D reconstruction systems. When given a set of RGB images as input, a classic 3D pipeline [34, 53] first identifies *co-visible* 3D points across images via pixel-wise visual features, and then recovers the spatial relationships among cameras. Such a "golden standard" framework has experienced huge success in practice and has given birth to numerous applications in robotics, AR, VR, etc. The reliance on correspondences, however, makes one ponder: what if the input images have little or no overlap? Does this still work when there are barely any co-visible 3D points in the scene (see Fig. 1)?

At first thought the answer is no. Predominant correspondence estimators focus on finding pixel pairs that describe the same, co-visible 3D points in the scene by matching their visual features. If the viewpoint differences across images are extreme, the pixels will be inherently different and cannot be matched, rendering current 3D systems to fail catastrophically. In contrast, humans can identify where the two photographs were taken with respect to the scene despite the large viewpoint variations. Such a remarkable capability comes from our prior knowledge of the underlying geometry, which helps us match pixels between images even if their exact correspondences are occluded or invisible in the other image. For instance, we know how the front and back of a human body should look like. Therefore, if we see a human face in one image and the back of a head in the other, we can easily associate them and infer that the two cameras are roughly 180 degrees apart. The aim of this paper is to equip 3D systems with similar abilities.

Towards this goal, we first ask the following question:

do we have to rely on pixels describing the same 3D points to recover camera poses[2]? While such (implicit) premises seem to lay the foundation for existing 3D reconstruction algorithms, as we will show in Sec. 3, the answer is negative. Our key observation is that epipolar geometry holds for *arbitrary pixels whose camera rays intersect in 3D*. Therefore, so long as one can identify those pixels, one can leverage them to recover relative camera poses, *regardless of whether the pixels are semantically or visually similar or not*. This interpretation is particularly exciting, as it allows one to go beyond the image space and establish geometric relationships among pixels even from extreme viewpoints.

Unfortunately, determining whether two camera rays intersect in 3D often requires camera poses to be known a priori, making the whole process a chicken-and-egg problem. Our key idea is to exploit prior knowledge of the foreground objects within the scene to break the loop. Specifically, we make use of *humans*, arguably one of the most common, salient "objects" in images. Consider the images in Fig. 1. If the system has prior knowledge about human shape and pose, it will know that a ray shooting through the human back in the leftmost image will intersect with the chest region on its way out. Furthermore, the intersecting chest pixel can be observed in the other image. Thus, we can find a pair of pixels that correspond to two intersecting camera rays with ease. Note that different from classic correspondences, these two pixels do not depict the same 3D point and thus cannot be found via visual similarities. Since we establish the geometric connection virtually by hallucinating a 3D shape, we call them *virtual correspondences* (VCs).

With this inspiration in mind, we first define virtual correspondences and present a methodology to derive them from images containing humans. We then showcase how VCs can be seamlessly integrated with the classic bundle adjustment algorithm, resulting in a generalized structure from motion (S*f*M) framework that could be applied to both traditional setup and extreme-view scenarios. We evaluate the effectiveness of our approach on the CMU Panoptic dataset [40, 42], the Mannequin Challenge dataset [51], and multiple challenging in-the-wild images. Our method significantly outperforms prior art in challenging extreme-view scenarios and is comparable in the conventional, densely overlapping setup. Importantly, our estimated poses from extreme viewpoints unleash the potential of multiple downstream applications such as scene reconstruction from multi-view stereo and novel view synthesis in challenging scenarios.

In summary, we make the following contributions:

1. We present virtual correspondences, a novel concept for 3D reconstruction algorithms, and establish its geometric connection to existing correspondences.
2. We develop a method to estimate VCs from images with humans and showcase how to integrate them into existing 3D frameworks. The new framework can be applied to a wide range of scenarios while also reduces to the classical S*f*M when no VCs are found.
3. We exploit the estimated camera poses for multiple downstream tasks and empirically show that our method enables extreme-view scenarios that were not feasible.

[2]We will ignore other primites such as lines or planes for now.

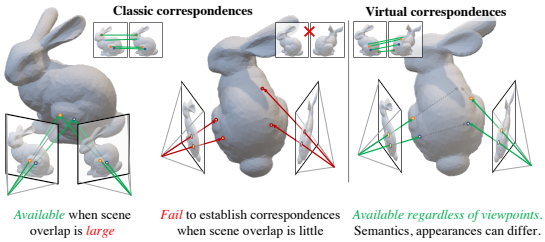


Figure 2. **Classic correspondences vs. virtual correspondences.**

2. Related Work

Correspondences: Correspondence estimation aims to identify pixels that are projections of the same 3D point across multiple images [34, 56]. The task has been the cornerstone of various computer vision problems for decades, since the pixel-level association allows one to recover the structure and motion of the world effectively [8, 35, 53, 67]. Prevalent approaches focus on hand-crafted [7, 13, 50, 55, 65, 77] or learned [20, 23, 60, 81–84] robust *visual* features that can distinguish one pixel from the others in diverse scenarios. While impressive performance has been achieved [66, 74], these methods fall short when there is little overlap among input images, as there are hardly any co-visible 3D points. Semantic correspondence estimation [16, 32, 33, 37, 46, 94, 95], on the other hand, focuses on detecting pixels with specific semantics (*e.g.*, human facial keypoints). With the help of domain knowledge, they are usually more robust to variations in viewpoint, appearance, and sometimes even occlusions [14, 36]. Unfortunately, they still require a set of semantic keypoints to be co-visible across multi-view images to enable 3D reconstruction. In contrast, our novel virtual correspondences do not have these constraints. VCs can be the projection of different 3D points and can have completely different appearances and semantics (*e.g.*, chest pixel *v.s.* back pixel). This allows us to establish geometric relationships among pixels even when the input images have no co-visible 3D points.

Extreme pose estimation: There has been a surge of interest in estimating relative 3D poses among a set of little- or non-overlapping RGB(D) images [12, 39, 64, 69, 88]. Different from the classical small- or wide-baseline setup, the viewpoint variation of this task is much more extreme where few co-visible regions can be found, rendering traditional matching-based approaches unsuitable. To address this challenge, researchers have proposed to either directly predict the transformation with deep neural nets [12, 15], or adopt the hallucinate-then-match paradigm [6, 29, 64, 88, 90]. Our work

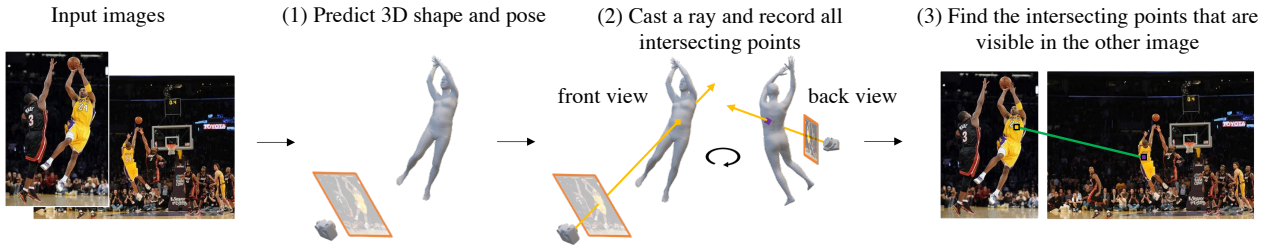


Figure 3. **Pipeline.** We first predict the 3D shape and pose of the basketball player from the left image. Then we cast a ray and record all the points it hits, *i.e*. the belly button and his back. While the two images barely overlap, the right image does observe the back of player. We can thus tell that the rays of the two pixels intersect at 3D and are virtual correspondences. We conduct the same process for the right image too.

lies under the broad umbrella of the hallucination paradigm, as we derive virtual correspondences from hallucinated human shape priors and combine them with epipolar geometry. We adopt a pixel-level correspondence representation, which seamlessly integrates with prevailing 3D reconstruction algorithms and can be naturally extended to the multi-view setup. In contrast, previous methods only consider two frames at a time [39, 64, 88, 90], as the customized matching and optimization step prohibits them from scaling up easily.

Structure from motion (S*f*M): Given a set of images, the goal of S*f*M algorithms [2, 8, 17, 27, 62, 76, 78, 79] is to recover both the camera poses and the (sparse) 3D geometry of the scene. Prevailing S*f*M systems [68, 72, 73, 85] have enjoyed great success when the images are captured densely with large overlaps, yet they suffer drastically when the input views are sparse and have little overlap. To alleviate this issue, researchers have sought to exploit motion patterns [4, 5, 75] or semantic keypoints of the objects [22, 86] to aid the reconstruction. However, they require sequences of frames as input (with static cameras) or the same set of keypoints to be visible across all views, which largely limits their applicability. Our virtual correspondences, in comparison, are much more flexible: while our VCs are also derived from objects, specifically humans, the corresponding pixels can have completely different semantics and appearances. This allows us to establish matches even if the input images have no co-visible 3D points. Our approach also shares similar insights with non-rigid S*f*M algorithms, which leverage shape dictionaries (*i.e*., priors) to constrain the solution space [3, 11, 18, 19, 38, 48, 80]. However, unlike these approaches, we do not require 2D correspondences to be given a priori. We instead exploit shape priors to establish VCs across views that conventionally do not have correspondences. As we will show in the experimental section, VCs open the door to a range of possibilities and broaden the applicable domain of S*f*M.

3D human estimation: Our work is also related to 3D human reconstruction approaches [31, 87, 89]. With the flourishing of deep learning, these methods have made tremendous progress, either from a single image [41, 44, 47] or multi-view images [21, 22, 25, 61]. While these approaches mostly focus on the quality of the reconstructed shape, we attempt to recover accurate camera poses with human shape priors. More recently, researchers have exploited human keypoints to refine camera poses [22, 63], but by virtue of VCs, our method is more flexible and does not require the same keypoints to be co-visible across views. As we will show in Sec. 3.3, our bundle adjustment formulation is a superset of theirs. Our work also shares similar insights with human silhouette matching [70, 71], since we both do not rely on appearance matching to establish correspondences, allowing us to generalize to extreme-view setting. However, there exist several differences: First, while they require video sequences to constrain the solution space, a single image pair suffice for us. Second, they capitalize on sufficient motion of the object over the space for matching, whereas we exploit deep shape priors to estimate the correspondences. Third, their frontier points are still co-visible across cameras, yet our VCs may correspond to completely different 3D points.

3. Approach

Our aim is to equip existing 3D systems with the ability to reason and associate images geometrically even if they have little or no overlap. We seek to devise a method that can be seamlessly integrated with existing 3D reconstruction frameworks such that the new model can be applied to both the conventional setup and the extreme setting. Towards this goal, we introduce a novel concept dubbed as *virtual correspondence (VC)*. VCs refer to a pair of pixels whose camera rays intersect in 3D. However, unlike classic correspondences, they do not need to describe the same 3D points, and can have completely different semantics and appearances. This makes VCs much more flexible and allows VCs to be established even when there is little overlap among images. Importantly, VCs conform to epipolar geometry and can be combined with prevailing 3D systems naturally. We unfold this section by formally defining VCs and discussing their relationships with existing correspondences. Then we present a method to estimate VCs through the lens of human shape priors. Finally we incorporate VCs into current S*f*M formulations, resulting in a framework that is much more general. For simplicity, we assume there are only two cameras, but the concepts and the method can be trivially

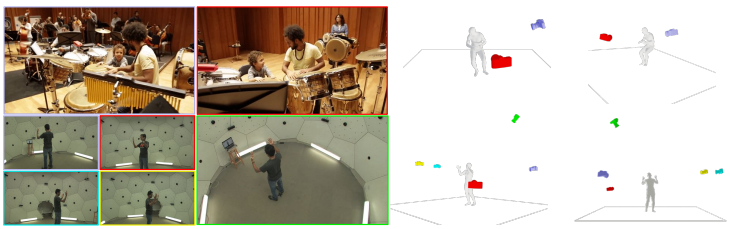


Figure 4. **Qualitative results.** (Left) Input images. (Right) Recovered camera poses. Human meshes are for illustration purpose.

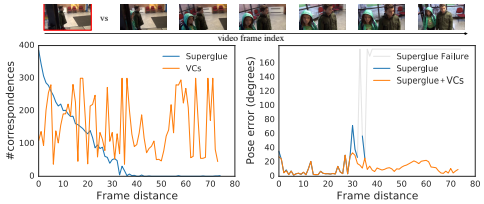


Figure 5. **Effects of camera distance.** We show the number of correspondences (left) and pose error (right) with increasing camera baseline.

extended to the multi-camera setup (as shown in Sec. 4).

3.1. Virtual Correspondences (VCs)

We first define virtual correspondences. Let $\mathcal{I}_1, \mathcal{I}_2 \in \mathbb{R}^{H\times W\times 3}$ be the images of the same scene captured at different viewpoints and $\boldsymbol{p}_1, \boldsymbol{p}_2 \in \mathbb{R}^2$ be the points in their respective image coordinates. Let $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{3\times 3}$ be the camera intrinsics and $[\mathbf{R}_1, \mathbf{t}_1], [\mathbf{R}_2, \mathbf{t}_2] \in \mathbb{R}^{3\times 4}$ be their extrinsic matrices. The ray marching from the camera center $\mathbf{o} \in \mathbb{R}^3$ through $\boldsymbol{p}$ can be written as $\boldsymbol{r}_{\boldsymbol{p}}(d) = \mathbf{R}^T(d\mathbf{K}^{-1}\bar{\boldsymbol{p}} - \mathbf{t})$, where $d > 0$ indicates the depth along the ray and $\bar{\cdot}$ refers to the homogeneous coordinate.

We say a point $\boldsymbol{p}_1$ in the first image and a point $\boldsymbol{p}_2$ in the second image, $(\boldsymbol{p}_1, \boldsymbol{p}_2)$, are *virtual correspondences* if there exists a pair of d's such that:

$$\boldsymbol{r}_{\boldsymbol{p}_1}(d_1) = \boldsymbol{r}_{\boldsymbol{p}_2}(d_2). \tag{1}$$

Since there is no constraint on where the intersection should happen, the rays can intersect at (i) co-visible 3D points, (ii) 3D points that are only visible in one image (and occluded in other other), or even (iii) invisible points (*e.g.*, free space, occupancy space, or points from occluded scene/objects).

The first scenario is exactly the definition of classic correspondences [20, 66]. The third scenario covers many cases in semantic correspondence where the target 3D points is invisible. For instance, researchers have exploited 2D human keypoints to reconstruct 3D joints [15, 22]. 3D joints, strictly speaking, lie within the human body and are not visible in images. VCs can therefore be seen as a generalization of multiple types of existing correspondences.

In the second and third scenario, VCs correspond to different 3D points in the scene. VCs can thus have different appearances and semantics, and even describe completely different parts of the scene. We show an example in Fig. 2 (right) where the pixels in the left image observe the leg while their VCs in the right see the back of the bunny. We refer the readers to supp. material for more illustrations.

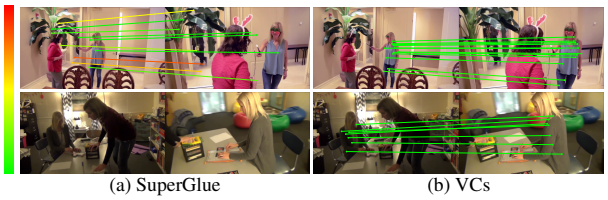
Another key property of VCs is that they conform to epipolar constraints — the two intersecting rays form an epipolar plane on which the VCs and camera origins lie. This allows us to exploit classic geometric algorithms to establish connections among non-overlapping images, greatly expanding the applicable domains of existing 3D algorithms. For instance, we cannot employ the five-point algorithm [53] for non-overlapping images in the past, since no correspondences exist. VCs, however, are more flexible and are not restricted to describing the same co-visible scene points. We can thus estimate VCs among the images and then solve for the essential matrix. We refer the readers to supp. material for more discussion on VCs and epipolar geometry.

While VCs are powerful, estimating them purely from 2D images is far from trivial. Without knowing the relative camera poses, one cannot exploit Eq. 1 to verify if two camera rays intersect. Furthermore, VCs may have completely different appearances and semantics, prohibiting us from employing similar approaches as classic correspondence estimators. Fortunately, there are many objects in the scene whose shapes we are familiar with. With such prior knowledge, we can hallucinate the shape of an object and estimate which part of the object a ray would intersect with on the other side. All one needs to do is then to find the rays (pixels) in other images that hit (see) the same intersecting point.

3.2. Exploiting Humans for VC Estimation

Based on the intuition above, we propose an approach to exploit shape priors for virtual correspondence estimation. We focus on humans, the most common "objects" in images.

Given a 2D image, we first exploit a deep network [41] to predict the 3D shape and pose of each person in the scene, as well as their relative poses to the camera. We use SMPL [54] as our representation since it allows us to reconstruct complete human mesh from partial observations. Then we cast a ray through each pixel and record all the 3D points where the rays intersect with the human mesh via ray-plane intersection (see Fig.3-mid). Finally, we identify if any of those 3D points are visible in other images by 2D-3D association. If there is, we say the two pixel rays intersect in 3D and the corresponding two pixels are VCs. Specifically, we use DensePose [30] to associate each pixel with each point on human mesh. If a ray hits the back of the mesh and DensePose tells us a pixel corresponds to the back, then these two pixels are VCs. Fig. 3 illustrates the process, which we repeat for all images. We note that our formulation is generic and can be potentially applied to other objects so long as there exist proper shape priors and surface mapping. We show an example on cars in Sec. 4.



(a) SuperGlue (b) VCs

Figure 6. **Qualitative comparison.** Classic correspondence estimators fail when images have little overlap, since there are no co-visible 3D points. VCs can be found in both scenarios so long as the camera rays intersect. The color indicates epipolar error.

3.3. Generalized Bundle Adjustment (BA)

Once we establish virtual correspondences, the next step is to jointly refine the camera poses as well as the sparse 3D scene geometry. Similar to classic S*f*M, we initialize the camera poses using RANSAC with the five-point algorithm [34] in the loop[3]. But instead of employing classic correspondences, we use VCs.

Since VCs could correspond to different 3D points (see Fig. 2), traditional triangulation approach cannot recover both 3D points. We thus leverage the initial shape estimation (predicted by deep nets) to compute the ray-surface intersection and record the first hits for each VC. The 3D points are then registered into the global coordinate system using the estimated camera poses from the five-point algorithm.

Since the estimated structure (*i.e.* the sparse 3D points) and poses depend heavily on the predicted shape priors, they may be noisy. We further refine the estimates by minimizing the distance between reprojected points and VCs. Formally, let $(\mathbf{X}^{j_1}, \mathbf{X}^{j_2})$ be the j-th pair of reconstructed 3D points and $(\boldsymbol{p}_{i_1}, \boldsymbol{p}_{i_2})$ be the associated VC pair from camera i_1 and camera i_2. Denote $\alpha = (i_1, i_2, j_1, j_2)$ as a tuple of corresponding indices. Our goal is to minimize:

$$\min_{\mathbf{R}_i, \mathbf{t}_i, \mathbf{X}^{j_1}, \mathbf{X}^{j_2}} \sum_{\alpha} \|\boldsymbol{p}_{i_1} - \pi_{i_1}(\mathbf{X}^{j_1})\|^2 + \|\boldsymbol{p}_{i_2} - \pi_{i_2}(\mathbf{X}^{j_2})\|^2$$
$$\text{s.t.} \quad \Big((\mathbf{X}^{j_1} - \mathbf{o}_{i_1}) \times (\mathbf{X}^{j_2} - \mathbf{o}_{i_2})\Big)^T (\mathbf{o}_{i_2} - \mathbf{o}_{i_1}) = 0, \tag{2}$$

where $\pi_i(\mathbf{X}) \sim \mathbf{K}_i(\mathbf{R}_i\mathbf{X} + \mathbf{t}_i)$ is the perspective projection operator, and the constraint enforces the two camera rays to be co-planar such that epi-polar geometry holds.

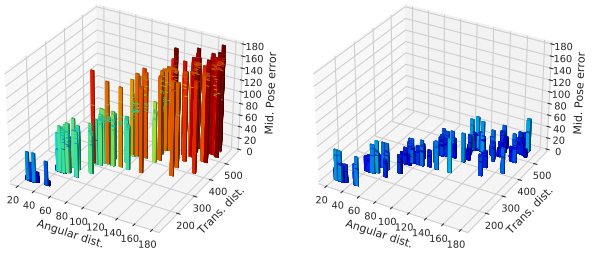
Using the constraint, we can further re-write one VC point as a function of the other:

$$\mathbf{X}^{j_2} = \mathbf{X}^{j_1} + a^j \cdot (\mathbf{X}^{j_1} - \mathbf{o}_{i_1}) + b^j \cdot (\mathbf{o}_{i_2} - \mathbf{o}_{i_1}). \tag{3}$$

The two free parameters a^j and b^j can be thought of as the "thickness" of the shape between the intersecting points. When both parameters become 0, the two 3D points merge into one, and VCs reduce to classic correspondences.

By replacing Eq. 3 into Eq. 2, we obtain an unconstrained minimization problem that is similar to, yet more generic than classic BA. Instead of refining a set of *co-visible 3D points*, we now adjust a bundle of *point tuples*. We, however, note that classic correspondences extracted with conventional methods such as SuperGlue [66] can still fit into this formulation by fixing $a^j = b^j = 0$. We use L-BFGS [58] to solve this non-linear least square problem. In practice, we treat Eq. 3 as a soft constraint since it works slightly better. We refer the readers to supp. material for more discussions.

[3]We assume the intrinsics are known or already estimated.



(a) Classic S*f*M (b) Our S*f*M

Figure 7. **Pose error vs. ground-truth pose distance.** The median pose error in classic S*f*M (left) increases with increasing camera baseline, while the median pose error for our method (right) stay low regardless of viewpoint differences.

Discussion: VCs can be combined with classic correspondences to improve the overall robustness and performance of 3D reconstruction systems (see Sec. 4). When the images barely overlap and few classic correspondences are available, the system can rely on VCs to recover the world and camera geometry. When the images do overlap, VCs can serve as additional visual cues and regularizers. VCs thus significantly expand the applicable setting of existing SfM systems.

4. Experiments

In this section, we first evaluate the effectiveness of virtual correspondence and our 3D system on two challenging datasets. Then we comprehensively study the characteristics of our method. With the estimated camera poses, we further conduct two downstream tasks, namely scene reconstruction with multi-view stereo and novel view synthesis, in difficult extreme-view cases. Finally, to showcase our method generalizes beyond human-based images, we demonstrate proof-of-concept results with cars.

4.1. Datasets

CMU Panoptic dataset: CMU Panoptic dataset [40, 42] is a large-scale, multi-view video dataset designed for human analysis. It provides ground-truth camera poses as well as person associations across views. The sequences were captured in a studio with (approximately) synchronized cameras widely spread across the dome, providing us a diverse set of viewpoints that are barely available in the real world (*e.g.*, cameras looking at a person from the top). We select 43 sequences from `pose`, `haggling`, and `dancing`. Each sequence contains 1∼3 people performing different actions. We divide the data into two splits. Each split comprises a

Pose estimation AUC (↑)	CMU Panoptic Studio			Mannequin Challenge		
Methods	@15°	@30°	@45°	@15°	@30°	@45°
SuperGlue [66]	10.02	16.74	19.36	26.38	34.85	39.10
LoFTR [74]	5.12	10.47	13.07	27.47	35.98	40.10
SIFT [55] + BA [68]	7.68	11.39	13.33	14.17	20.24	24.25
SuperPoint [20] + BA [68]	9.22	13.77	15.85	17.12	23.48	26.81
SuperGlue [66] + BA [68]	10.68	16.57	18.92	26.24	35.12	39.46
LoFTR [74] + BA [68]	8.35	14.52	17.01	27.51	36.32	40.55
Deep regression [49]	14.36	18.60	23.18	4.61	11.23	16.44
Deep optimization [9, 45]	7.88	27.17	42.42	15.38	47.08	63.67
Our S*f*M	**18.21**	**46.05**	**62.08**	**36.24**	**61.38**	**73.20**

Table 1. **Two frame relative pose estimation** on the CMU dataset and the MC dataset. First two rows perform five-point algorithm to derive camera poses. BA = Bundle Adjustment.

set of unique sequences and cameras without any overlap. Due to image quality, we only consider the videos captured by HD cameras. We sample a frame every five seconds to avoid similar human poses. We also run human detection on each sampled frame. If no person is present in the scene, we discard the frame. In total, we obtain 2955 image sets for each split, with each set containing 15-16 camera views. We refer the readers to the supp. material for more details.

Mannequin Challenge: Mannequin Challenge (MC) [51] is a dataset of internet video clips where the participants stay still in different poses, while the video-takers move freely in space and capture the event. These videos, by design, allow us to look at a static scene from various angles. We follow a similar pipeline as [51] to reconstruct the ground-truth camera trajectories and filter out snippets with small shifts in viewpoints or view directions. In the end, we obtain 18 video snippets where the cameras rotate by at least 90° within each sequence. To further increase pose diversity, we additionally collect 6 MC videos ourselves. Compared with the CMU dataset, the camera poses in MC videos are rather *generic* [28], yet the background scenes, which consist of both indoor and outdoor environments, are much more diverse. Finally, for each snippet, we compute the pose difference between each frame and the first frame. We sample a frame at every 20 percentile of the snippet and obtain ∼200 image pairs. All the images are treated as the test set.

4.2. Experimental Details

Metrics: Following previous work [10, 66, 91, 92], we employ the area under the cumulative error curve (AUC) to evaluate the recovered camera poses. We report the AUC at three different thresholds (15°, 30°, and 45°). The pose error is defined as the maximum of 1) the angular difference between predicted and GT rotation vectors; and 2) the angular difference between predicted and GT translation vectors. We report angular difference for translation since it can only be recovered up to a scaling factor [34]. As for 3D reconstruction, there is no standard protocol to compare point clouds directly produced by S*f*M systems because each S*f*M algorithm can choose which 3D points to reconstruct.

Initialization		BA		Pose estimation AUC		
SG	VCs	SG	VCs	@15°	@30°	@45°
✓	-	-	-	10.02	16.74	19.36
✓	✓	-	-	10.29	31.27	48.96
✓	-	✓	-	10.68	16.57	18.92
-	✓	-	✓	15.89	43.92	60.38
✓	✓	✓	✓	**18.21**	**46.05**	**62.08**

Table 2. **Ablation study** on the CMU dataset. SG = SuperGlue.

Furthermore, there is no ground-truth shape for both datasets. We thus follow [43] to compute the silhouette accuracy between the rendered mask and the 2D segmentation mask.

Baselines: We compare our method against a wide range of relative pose estimation methods. For traditional matching-based methods, we first detect the key points and extract their corresponding features with SIFT [55] or SuperPoint [20]. We then establish classic correspondences with either nearest neighbour matching with ratio test [55] or SuperGlue (SG) [66]. We also compare with LoFTR [74]. We further use RANSAC [26] coupled with the five-point algorithm to filter outliers. We then incrementally recover and bundle adjust the image poses with COLMAP [68]. Alternatively, if there are only two views, we also perform pose estimation with the five-point algorithm and essential matrix decomposition. Next, for deep regression methods, we employ a state-of-the-art pose estimation network [24] to predict the relative camera pose between an image pair. Finally, we compare against a deep optimization approach that estimates camera poses by aligning 3D shapes. The baseline is inspired by the state-of-the-art indoor extreme pose estimation method [64] and can be seen as a variant for humans. Specifically, we utilize the latest EFT-Net [41] to reconstruct 3D human models and align them with ICP [9]. To avoid local minima, we first register the shapes based on their canonical coordinates. Next, we associate each part of the shape based on its semantics. We further prune out the limbs and exploit only torso and head during matching since these two parts are more robust in practice. These strategies drastically improve the performance of this baseline.

Implementation details: Our 3D system considers both classic correspondences and VCs. We exploit SuperGlue [66] to estimate classic correspondences and ReID-Net [93] to match a person across multiple viewpoints. For the deep regression baseline, we train and validate on the training split of CMU dataset. For the rest of the learning based approaches, *including our method*, we adopt the pre-trained weights provided by the authors and conduct inference only.

4.3. Experimental Results

CMU Panoptic Studio: As shown in Tab. 1(left), our S*f*M outperforms all baselines at all thresholds in the two-frame pose estimation task. SuperGlue [66] ranks second

Figure 9. **Reconstructed mesh** using our method + multi-view stereo for two non-overlapping video sequences.

when the pose error threshold is low, but deep optimization [9, 44] surpasses it when the threshold increases. This is expected since matching-based approaches can produce accurate estimation when classic correspondences are available, yet fail catastrophically when the viewpoints are very different. Deep optimization, in contrast, is not as accurate when the view difference is small, but has fewer fatal failures. Our approach, which exploits both classic and virtual correspondences, does not suffer from either catastrophic wide-baseline failures or inaccurate narrow-baseline matching. Our S*f*M has a median error of 15.7° and the pose error at the 80th percentile is less than 24°. In contrast, the median error of deep optimization is 23.5° and the pose error at the 80th percentile is 44°. Compared to EFT-Net, we improve the silhouette accuracy from 74% to 81%.

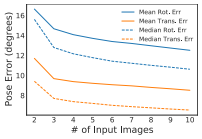


Figure 8. **Error vs. # of images.**

We also investigate how our S*f*M scales with more input images. Following COLMAP [68], we start from an image pair and then incrementally register new images. As shown in Fig. 8, the pose error reduces as more images are added. The reduction is most significant when registering the third image. We hypothesize this is because the third image greatly increases the overlap among the images, providing more reliable classic correspondences during bundle adjustment. We also compare our approach with the classic S*f*M methods. Our AUC continuously outperforms the baselines at all thresholds (*e.g.*, @15°: 28.4 vs 17.6). We refer the readers to the supp. material for full ablation table, cumulative error plots, and detailed performance of all methods with respect to the input images.

Mannequin Challenge: As shown in Tab. 1(right), our method outperforms all baselines at all thresholds. Despite more diverse scenes, our AUC on the MC dataset is higher than that of the CMU dataset. We hypothesize this is because the viewpoint changes in the MC dataset are less significant than the CMU dataset, due to how the dataset was collected.

Qualitative results: We showcase our results on a two-view MC image pair, and a five-view CMU image set in Fig. 4. Our testing scenarios are typically very challenging, with large view variations and small proportion of co-visible regions. Nevertheless, our proposed S*f*M framework is able to recover both relative poses as well as the parametric human shape accurately.

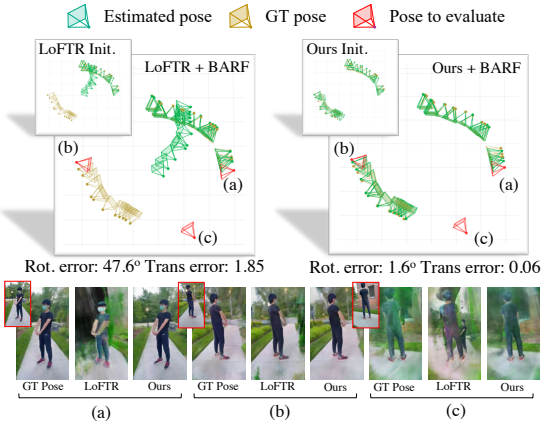


Figure 10. **Novel view synthesis:** (top left) camera poses initialized with LoFTR and refined by BARF; (top right) camera poses initialized with our method and refined by BARF; (bottom) images synthesized at novel views by BARF initialized with GT pose, LoFTR and our framework respectively.

4.4. Analysis

Ablation study: To gain more insights into the contribution of each component, we evaluate our method with different configurations on the CMU dataset. As shown in Tab. 2, by simply exploiting VCs during initialization, our method surpasses classic S*f*M in terms of AUC at large thresholds. Additionally, the ablation study shows that bundle adjustment is critical for VCs. We conjecture this is because VCs are constructed from initial shape priors, which are noisy. By bundle adjusting the line segments, we are essentially conducting maximum likelihood estimation [34] under Gaussian noise assumption on VC re-projection errors, which mitigates errors introduced by inaccurate VC pairs.

Effects of viewpoint changes: We use the MC dataset to illustrate how classic and virtual correspondences evolve with viewpoint changes and how it affects pose estimation. In general, the ground-truth camera pose difference is proportional to the video frame index distance. For each video, we compute the classic correspondences and VCs between all frames and the first frame, and then estimate the relative camera poses based on them. Since the number of classic correspondences decreases drastically when the viewpoint changes, classic S*f*M fails. In contrast, our S*f*M framework incorporates both classic correspondences and VCs to avoid failures. Fig. 5 shows an example of how our system produces decent estimation across all distances. We also showcase a "discrete" evaluation on the CMU dataset in Fig. 7. The pose error of classic S*f*M methods increases significantly with respect to the ground-truth camera pose distances (the diagonal direction), while our S*f*M performs

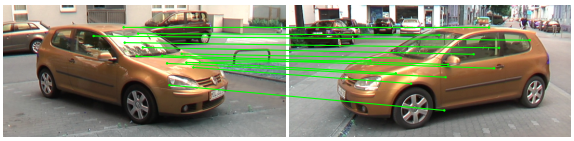


Figure 11. **Virtual correspondences from cars.**

consistently across all settings.

Reliability of human parts: We compute the histogram over all VCs on both datasets. Around half of the VCs lie on human torso, and around 12% of VCs are derived from the human head. The remaining VCs uniformly spread across the whole body. Unlike the deep optimization baseline, we do not encode any prior knowledge into our system, yet our approach is able to automatically discover that human torso is the most reliable parts within the predicted 3D shapes. Fig. 6 shows a subset of VCs selected by our S*f*M system.

Generalization to in-the-wild images: Our approach can be applied to real-world image collections without bells and whistles. We test our system on a pair of movie frames and two pairs of sports photos in Fig. 1. Even though the cameras are far apart and the images are slightly asynchronous, our system still produces reasonable estimates. More results on classic movies and sports events can be found in supp. materials.

Limitations: Our approach relies heavily on the predicted shape priors. While we can handle noisy predictions by pruning out the outlier VCs during geometric verification, if the initial estimation is completely wrong (which our system can detect by comparing silhouette consistency, DensePose consistency, *etc.*), we will not be able to construct VCs. Additionally, similar to classic S*f*M algorithms, we assume the scene is static. While we can tolerate slight movements (see Fig. 1), it fails when human poses change significantly.

4.5. Applications

Scene reconstruction with multi-view stereo (MVS): We first show how VCs enable coherent multi-view 3D reconstruction from non-overlapping videos. This type of capture is common in practice, yet most S*f*M and MVS systems can only handle each sequence individually, resulting in two disjoint 3D reconstructions. Our VCs, in contrast, are able to recover relative poses even from non-overlapping images, which allows us to obtain a single, coherent MVS point cloud to unlock further geometry processing such as mesh reconstruction. We use RealityCapture [1] to reconstruct the 3D scene by initializing the camera poses with our estimation. The resulting high-quality meshes suggest that the recovered camera poses and the extracted point clouds are accurate (see Fig. 9). In contrast, both RealityCapture's built-in 3D reconstruction pipeline and COLMAP [68] fail due to non-overlapping viewpoints.

Novel view synthesis: We further demonstrate the effectiveness of our approach in extreme-view scenarios through the task of novel view synthesis, which relies heavily on input poses. In particular, we adopt BARF [52], an approach that can learn a neural radiance field [57] and refine camera poses simultaneously. We again use two non-overlapping video sequences. We initialize BARF with the poses recovered by our method and LoFTR [74] respectively. Our estimated poses, which are already fairly accurate, are further refined through the course of BARF training (see Fig. 10). In contrast, LoFTR [74] fails to estimate the relative camera poses among the two sequences correctly (see the green cameras) and the resulting BARF training gets stuck in local minima. We also evaluate the learned radiance field with novel, extrapolated poses. Our view syntheses results are comparable to those trained with GT poses. As expected, the quality degrades when the evaluation pose deviates too far from the training poses, especially for the background scene that is unseen in the training videos. However, we can still see a person standing on the pavement and observe the structure of the scene. On the other hand, due to the incorrect LoFTR poses, the baseline fails to produce realistic results.

Extending VCs to other objects: As a proof-of-concept, we exploit canonical 3D deformable mapping [59] as shape priors and adapt our method to cars. As shown in Fig. 11, we are able to estimate VCs and recover relative poses effectively (pose error: 16°) even from extreme viewpoints. We refer the readers to supp. material for more details.

5. Conclusion

In this paper, we introduced a novel concept called virtual correspondences – a pair of image points whose camera rays intersect in 3D. Unlike classic correspondences, virtual correspondences do not need to describe the same, co-visible 3D point and thus are not constrained by visual or semantic similarities, making it possible to match images with little or no overlap. We proposed a method to extract virtual correspondences based on prior knowledge of foreground objects in the image, and integrate with existing 3D frameworks. Our experiments on two challenging human-based datasets show that virtual correspondences are critical towards successful camera pose estimation and downstream multi-view stereo and novel view synthesis in extreme-view scenarios.

Social impact: Our method alleviates the need to capture dense views for camera pose estimation and 3D reconstruction, and has the potential to reduce storage and computational costs. Unfortunately, it could also be exploited by surveillance and may raise privacy concerns as 3D reconstruction from few images becomes more accessible.

References

[1] Explore the possibilities of realitycapture, 2021. 8

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *ACM Communications*, 2011. 3

[3] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *NIPS*, 2008. 3

[4] Roland Angst and Marc Pollefeys. Static multi-camera factorization using rigid motion. In *ICCV*, 2009. 3

[5] Roland Angst and Marc Pollefeys. Multilinear factorizations for multi-camera rigid structure from motion problems. *IJCV*, 2013. 3

[6] Mohamed El Banani, Jason J Corso, and David F Fouhey. Novel object viewpoint estimation through reconstruction alignment. In *CVPR*, 2020. 2

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006. 2

[8] Paul Beardsley, Phil Torr, and Andrew Zisserman. 3d model acquisition from extended image sequences. In *ECCV*, 1996. 2, 3

[9] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, 1992. 6, 7

[10] Eric Brachmann and Carsten Rother. Neural-guided ransac: Learning where to sample model hypotheses. In *ICCV*, 2019. 6

[11] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2000. 3

[12] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *CVPR*, 2021. 2

[13] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In *ECCV*, 2010. 2

[14] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2019. 2

[15] Kefan Chen, Noah Snavely, and Ameesh Makadia. Wide-baseline relative camera pose estimation with directional learning. In *CVPR*, 2021. 2, 4

[16] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. 2016. 2

[17] MAR Cooper and S Robson. Theory of close range photogrammetry. 1996. 3

[18] Alessio Del Bue. A factorization approach to structure from motion with shape priors. In *CVPR*, 2008. 3

[19] Alessio Del Bue, Xavier Llad, and Lourdes Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *CVPR*, 2006. 3

[20] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR*, 2018. 2, 4, 6

[21] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. In *TPAMI*, 2021. 3

[22] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *ECCV*, 2020. 3, 4

[23] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv*, 2019. 2

[24] Sovann En, Alexis Lechervy, and Frédéric Jurie. Rpnet: An end-to-end network for relative camera pose estimation. In *ECCVW*, 2018. 6

[25] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *CVPR*, 2021. 3

[26] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *ACM Communications*, 1981. 6

[27] Andrew W Fitzgibbon and Andrew Zisserman. Automatic camera recovery for closed or open image sequences. In *ECCV*, 1998. 3

[28] William T Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368(6471):542–545, 1994. 6

[29] Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Visual correspondence hallucination. *arXiv*, 2021. 2

[30] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 4

[31] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 3

[32] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *CVPR*, 2016. 2

[33] Kai Han, Rafael S Rezende, Bumsub Ham, Kwan-Yee K Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *ICCV*, 2017. 2

[34] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000. 1, 2, 5, 6, 7

[35] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 1981. 2

[36] Peiyun Hu and Deva Ramanan. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *CVPR*, 2016. 2

[37] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *CVPR*, 2019. 2

[38] Sebastian Hoppe Nesgaard Jensen, Mads Emil Brix Doest, Henrik Aanæs, and Alessio Del Bue. A benchmark and evaluation of non-rigid structure from motion. *IJCV*, 2021. 3

[39] Linyi Jin, Shengyi Qian, Andrew Owens, and David F Fouhey. Planar surface reconstruction from sparse views. 2021. 2, 3

[40] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 2, 5

[41] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2021. 3, 4, 6

[42] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 2017. 2, 5

[43] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 6

[44] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3, 7

[45] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 6

[46] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *CVPR*, 2017. 2

[47] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 3

[48] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *ICCV*, October 2019. 3

[49] Nilesh Kulkarni, Ishan Misra, Shubham Tulsiani, and Abhinav Gupta. 3d-relnet: Joint object and relational network for 3d prediction. In *ICCV*, 2019. 6

[50] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. Brisk: Binary robust invariant scalable keypoints. In *ICCV*, 2011. 2

[51] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 2, 6

[52] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. *arXiv*, 2021. 8

[53] H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 1981. 1, 2, 4

[54] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 4

[55] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 2004. 2, 6

[56] Yi Ma, Stefano Soatto, Jana Kosecka, and S Shankar Sastry. *An invitation to 3-d vision: from images to geometric models*. Springer Science & Business Media, 2012. 2

[57] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 8

[58] Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980. 5

[59] David Novotny, Roman Shapovalov, and Andrea Vedaldi. Canonical 3D Deformer Maps: Unifying parametric and non-parametric methods for dense weakly-supervised category reconstruction. In *NeurIPS*, 2020. 8

[60] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. *arXiv*, 2018. 2

[61] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. *arXiv*, 2020. 3

[62] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *IJCV*, 2004. 3

[63] Jens Puwein, Luca Ballan, Remo Ziegler, and Marc Pollefeys. Joint camera pose estimation and 3d human pose estimation in a multi-camera setup. In *ACCV*, 2014. 3

[64] Shengyi Qian, Linyi Jin, and David F Fouhey. Associative3d: Volumetric reconstruction from sparse views. In *ECCV*, 2020. 2, 3, 6

[65] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, pages 2564–2571, 2011. 2

[66] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2, 4, 5, 6

[67] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002. 2

[68] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3, 6, 7, 8

[69] Mohammad Amin Shabani, Weilian Song, Makoto Odamaki, Hirochika Fujiki, and Yasutaka Furukawa. Extreme structure from motion for indoor panoramas without visual overlaps. In *ICCV*, 2021. 2

[70] Sudipta N Sinha and Marc Pollefeys. Camera network calibration and synchronization from silhouettes in archived video. *IJCV*, 2010. 3

[71] Sudipta N Sinha, Marc Pollefeys, and Leonard McMillan. Camera network calibration from dynamic silhouettes. In *CVPR*, 2004. 3

[72] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *SIGGRAPH*. 2006. 3

[73] Noah Snavely, Steven M Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *IJCV*, 2008. 3

[74] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 2, 6, 8

[75] Chris Sweeney, Aleksander Holynski, Brian Curless, and Steve M Seitz. Structure from motion for panorama-style videos. *arXiv*, 2019. 3

[76] Richard Szeliski and Sing Bing Kang. Recovering 3d shape and motion from image streams using nonlinear least squares. *Journal of Visual Communication and Image Representation*, 1994. 3

[77] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *TPAMI*, 2009. 2

[78] Philip HS Torr and Andrew Zisserman. Feature based methods for structure and motion estimation. In *International workshop on vision algorithms*, 1999. 3

[79] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, 1999. 3

[80] Chaoyang Wang, Chen-Hsuan Lin, and Simon Lucey. Deep nrsfm++: Towards unsupervised 2d-3d lifting in the wild. In *3DV*, 2020. 3

[81] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *ECCV*, volume 12346, 2020. 2

[82] Shenlong Wang, Linjie Luo, Ning Zhang, and Jia Li. Autoscaler: scale-attention networks for visual correspondence. *arXiv*, 2016. 2

[83] Shenlong Wang, Sean Ryan Fanello, Christoph Rhemann, Shahram Izadi, and Pushmeet Kohli. The global patch collider. In *CVPR*, 2016. 2

[84] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *CVPR*, 2019. 2

[85] Changchang Wu et al. Visualsfm: A visual structure from motion system. 2011. 3

[86] Yan Xu, Yu-Jhe Li, Xinshuo Weng, and Kris Kitani. Wide-baseline multi-camera calibration using person re-identification. In *CVPR*, June 2021. 3

[87] Ze Yang, Siva Manivasagam, Ming Liang, Bin Yang, Wei-Chiu Ma, and Raquel Urtasun. Recovering and simulating pedestrians in the wild. In *CoRL*, 2020. 3

[88] Zhenpei Yang, Jeffrey Z Pan, Linjie Luo, Xiaowei Zhou, Kristen Grauman, and Qixing Huang. Extreme relative pose estimation for rgb-d scans via scene completion. In *CVPR*, 2019. 2, 3

[89] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *CVPR*, 2021. 3

[90] Zhenpei Yang, Siming Yan, and Qixing Huang. Extreme relative pose network under hybrid representations. In *CVPR*, 2020. 2, 3

[91] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. *arXiv*, 2018. 6

[92] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *ICCV*, 2019. 6

[93] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv*, 2019. 6

[94] Tinghui Zhou, Yong Jae Lee, Stella X Yu, and Alyosha A Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *CVPR*, 2015. 2

[95] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A Efros. Learning dense correspondence via 3d-guided cycle consistency. In *CVPR*, 2016. 2