Toward Planet-Wide Traffic Camera Calibration

Khiem Vuong Robert Tamburo Srinivasa G. Narasimhan
{kvuong,rtamburo,srinivas}@andrew.cmu.edu
Carnegie Mellon University

Figure 1. Our framework allows us to perform 3D scene reconstruction and precise localization of over 100 real-world traffic cameras distributed globally across multiple countries, with the potential to scale to any camera with sufficient street-level imagery. **(Left)**: Highlighting the reconstruction and localization of traffic cameras at specific chosen locations. **(Right)**: Demonstrating 7 cameras positioned within an urban intersection, accurately localized with respect to the reconstructed 3D scene. **Please zoom in for better visualization.**

Abstract

Despite the widespread deployment of outdoor cameras, their potential for automated analysis remains largely untapped due, in part, to calibration challenges. The absence of precise camera calibration data, including intrinsic and extrinsic parameters, hinders accurate real-world distance measurements from captured videos. To address this, we present a scalable framework that utilizes street-level imagery to reconstruct a metric 3D model, facilitating precise calibration of in-the-wild traffic cameras. Notably, our framework achieves 3D scene reconstruction and accurate localization of over 100 global traffic cameras and is scalable to any camera with sufficient street-level imagery. For evaluation, we introduce a dataset of 20 fully calibrated traffic cameras, demonstrating our method's significant enhancements over existing automatic calibration techniques. Furthermore, we highlight our approach's utility in traffic analysis by extracting insights via 3D vehicle reconstruction and speed measurement, thereby opening up the potential of using outdoor cameras for automated analysis. Code and dataset will be available on the project website.

1. Introduction

With the recent advances in vision techniques, traffic cameras have gained numerous applications, including vehicle speed measurement [19, 51], automated traffic analytics [40, 44], and accident/anomalies detection [7, 19, 48], just to name a few. To enable such applications, camera calibration is a crucial requirement. Camera calibration includes estimating both intrinsic parameters (focal length and distortion coefficients) and extrinsic parameters (orientation and position of the camera) in metric real-world coordinates. In addition, for many downstream applications, the metric geometry of the scene, including the ground plane is often necessary. However, such calibration information is not readily available in most cases.

Despite extensive literature on traffic camera calibration, existing approaches suffer from various limitations. Traditional methods, such as those relying on checkerboard-based calibration [53, 56], are not practically scalable since physical access to the scene is required. Other techniques require manual inputs, like identifying landmarks with known dimensions like road markings which can be time-consuming and subject to human error [10, 54]. Certain approaches rely on estimating or assuming specific pri-

ors, such as vanishing points [15, 28, 50], average vehicle size [12], or camera height [54], and can introduce inaccuracies with limited generalizability to novel scenarios.

To address these challenges, we introduce a novel framework for automatically acquiring accurate metric 3D scene reconstruction and calibration of stationary traffic cameras at real-world street intersections. To achieve this, we leverage the vast amount of high-quality, geo-referenced, and calibrated images available in Google Street View (GSV) [21]. By utilizing GSV, we construct a metric-scale 3D scene reconstruction near the desired traffic camera location. Given that GSV offers panorama images, we improve our 3D reconstruction by enforcing known relative poses among perspective images sampled from the same panorama. Next, we employ state-of-the-art (SOTA) camera localization techniques, leveraging recent advances in learned feature matching, such as SuperPoint [13] and SuperGlue [45], to establish robust 2D-3D correspondences. This enables us to infer the traffic camera's intrinsic and extrinsic parameters accurately. It is worth noting that while our emphasis in this paper is on the utilization of GSV, our framework is adaptable to any source of street-level imagery, including Bing's Streetside [8], Mapillary Maps [38], user-captured data from smartphones (with GPS information), and other similar sources. This aspect highlights the scalability and versatility of our method, further extending its potential to achieve planet-wide camera calibration.

We demonstrate 3D scene reconstruction and accurate localization at over 100 traffic cameras across multiple countries and continents, with the potential to generalize to any novel camera where sufficient nearby street-level imagery is available. For quantitative validation, we propose a new dataset containing 20 fully calibrated traffic cameras at diverse urban scenes under varying capture conditions. Through extensive quantitative and qualitative experiments, we demonstrate the significant improvements of our method over existing SOTA methods in both intrinsic and extrinsic calibration. Leveraging accurate calibration, we illustrate its capabilities in the domains of 3D reconstruction and velocity measurement for moving vehicles, thus facilitating the extraction of valuable insights from the data.

To summarize, our main contributions are:

- We develop a scalable framework utilizing street-level imagery to create precise 3D models for accurate global calibration of traffic cameras, with successful localization of over 100 cameras and potential for broader application.
- We introduce a novel dataset featuring 20 fully calibrated traffic cameras, capturing diverse urban scenes under varying conditions, serving as a valuable benchmark for future research.
- We demonstrate the framework's efficacy in traffic analysis through automated extraction of insights via 3D vehicle reconstruction and speed measurement.

2. Related Work

Camera calibration involves two key aspects: 1) *intrinsics calibration*, which takes into account perspective projection (focal length, principal points, etc.) and potentially corrects for radial and tangential distortion, and 2) *extrinsics calibration*, which refers to camera rotation and translation, usually defined with respect to the ground plane. Note that for practical real-world applications like speed measurement, the extrinsic parameters must be expressed in *metric* units, often referred to as the *metric scene scale*.

Generic Camera Calibration: Within the domain of camera calibration, several methods have been established. Two widely employed gold-standard techniques, presented by Zhang et al. [56] and Tsai et al. [53], utilize planar calibration targets to estimate intrinsic and extrinsic camera parameters. Although those methods achieve sub-pixel calibration accuracy, they prove infeasible for traffic cameras positioned in challenging and potentially inaccessible locations, thereby limiting their scalability and practicality. Additionally, despite the capability of learning-based methods [9, 34] to recover focal length and distortion parameters from a single image, they usually do not generalize well to out-of-distribution data such as traffic cameras.

Traffic Camera Calibration: In the context of traffic scene analysis, a review of available methods has been presented by Sochor et al. [51]. Some approaches [10, 22, 24] rely on detecting vanishing points at road marking intersections, utilizing vehicle motion to calibrate the camera [12, 15, 16, 46], or involving manual measurements of dimensions on the road plane [14, 29, 35–37, 41, 49]. Various techniques have also been proposed for estimating the scene scale. For example, [15] employed a 3D bounding box around vehicles and their average dimensions to compute the scale, and [50] suggested using the alignment of a 3D model and a bounding box for scale inference. However, it is important to note that these methods are not without limitations, particularly in terms of scalability and accuracy. Manual techniques demand labor-intensive measurements of landmarks and dimensions. Meanwhile, automatic approaches relying on vehicle 3D model or vanishing points [4, 6, 16, 28, 50] still manifest notable errors and sensitivity to the quality of estimated geometric cues, especially when certain assumptions are compromised, e.g., non-straight vehicle motion, pronounced camera distortion, different viewpoints, or lack of knowledge of the exact make/model of vehicles, etc. On the other hand, our method does not make any assumptions about scene geometry or vehicle motion and instead takes advantage of the extensive collection of geo-registered panoramic street-level imagery, offering a novel, practical, and scalable solution for accu-

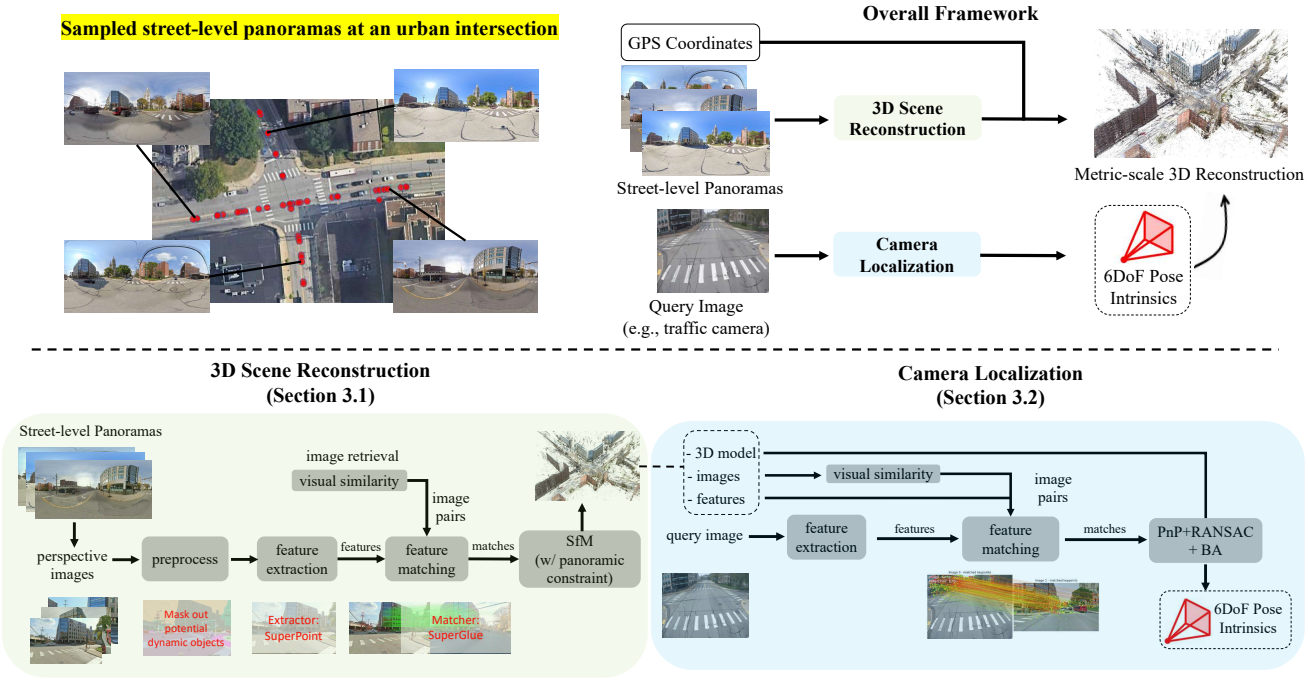


Figure 2. **Top**: Using street-level panoramas and GPS data from GSV, we reconstruct the scene in 3D for a metric-scale representation. With a query image from a traffic camera, we perform camera localization to determine intrinsic parameters and camera pose w.r.t. the 3D scene. **Bottom**: More details on *3D Scene Reconstruction* (left) and *Camera Localization* (right).

rate camera calibration.

Traffic Camera Applications: A comprehensive survey on *Monocular Visual Traffic Surveillance* has been conducted by Zhang et al. [55]. While applications like vehicle counting rely on 2D data, tasks demanding 3D insights such as speed estimation and distance measurement rely on precise camera calibration. In the context of speed measurement, recent methods [4,28,43] excel within specific scenarios but face limitations in generalizing to unfamiliar data. Our approach provides a simple yet effective solution, automating the acquisition of camera intrinsics, extrinsics, and metric scene geometry. This fosters the integration of 3D techniques into automated traffic analysis.

3. Method

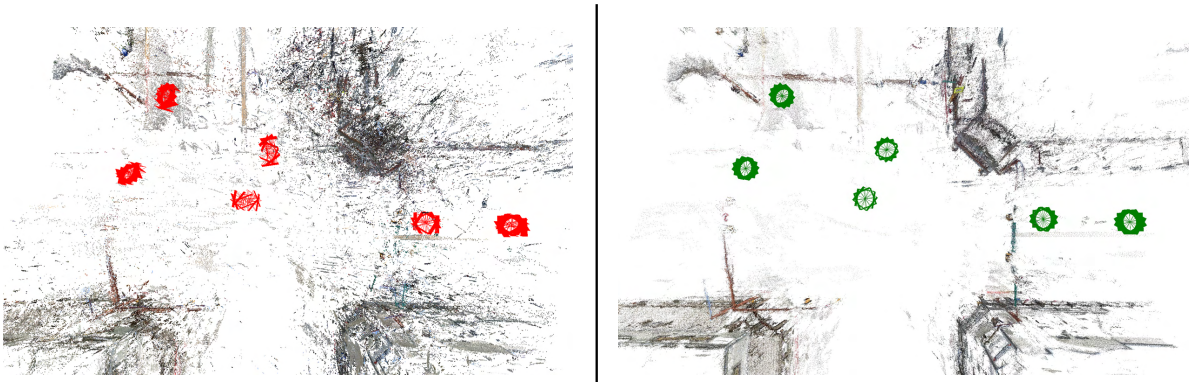
Our first objective is to construct a metric 3D reconstruction of the scene surrounding a chosen traffic camera's location, typically an intersection (Section 3.1). Following this, our goal is to localize the traffic camera within the reconstructed environment, thereby extracting both the intrinsic and extrinsic parameters of the camera (Section 3.2). The overall framework of our approach is depicted in Figure 2.

3.1. 3D Scene Reconstruction

To perform the reconstruction, we leverage Google Street View (GSV) [21] to build the scene's geometry around a specific GPS location. GSV is a street-level imagery database and a rich source of millions of panorama images with wide coverage all over the world (more than 10 million miles across 100 countries [20]). Every panorama image is geo-tagged with accurate GPS coordinates, capturing 360° horizontal and 180° vertical field-of-view (FoV) with high resolution (see top left of Figure 2). An overview of our **3D Scene Reconstruction** pipeline is illustrated in bottom left of Figure 2.

In particular, we first sample N panoramas (equirectangular frames) $\mathcal{E} = \{\mathcal{E}_i | i = 1 \dots N\}$ around the desired camera's location inside a radius of 40 meters. Since most components of a typical *structure-from-motion* (SfM) pipline [47] are primarily optimized for perspective images, we extract ideal, pinhole camera-style perspective projections from equirectangular images before performing 3D reconstruction. Specifically, from each equirectangular image $\mathcal{E}_i$, we extract T perspective images $\mathcal{I} = \{\mathcal{I}_{ij} | i = 1 \dots N, j = 1 \dots T\}$ that are uniformly sampled along the yaw direction with specified size and FoV, covering 360° horizontal FoV. Denoting $\mathbf{\Pi}(\cdot)$ as the projection function from equirectangular to perspective image, we can define each perspective image $\mathcal{I}_{ij}$ as:

$$\mathcal{I}_{ij} = \mathbf{\Pi}\Big(\mathcal{E}_i, \texttt{pitch=0}, \texttt{yaw=}\frac{2\pi * j}{T}, \texttt{fov=FOV}\Big) \quad (1)$$



Without Panoramic Constraint | **With Panoramic Constraint**

Figure 3. Enforcing known relative poses between perspective images from the same panorama leads to more accurate 3D reconstruction, especially in sparse-view scenarios. The recovered viewpoints in red (left) belonging to the same panorama do not coincide. **Please zoom in for better visualization.**

In practice, with image size of (1920×1080), we found the set of hyperparameters $\{\texttt{T=12}, \texttt{FOV=90°}\}$ to produce high-quality perspective images with sufficient overlap and minimal perspective distortions. Subsequently, we adapt COLMAP [47] to estimate camera pose $(R_{i,j}, t_{i,j})$ for each frame $\mathcal{I}_{ij}$.

Preprocessing. As dynamic objects often cause errors in the reconstruction, we apply a semantic segmentation method [11] to segment out potential dynamic objects such as vehicles and people and suppress feature extraction in these areas. For each perspective image (extracted from panorama image), the intrinsic camera matrix K_{ij} is known. Therefore, we fixed the shared camera intrinsics for all the frames during reconstruction.

Feature extraction and matching. We adopt SuperPoint [13] and SuperGlue [45] to establish correspondences among feature points across images. Instead of using exhaustive matching where each image is matched against every other image, we employ an adapted version of vocabulary tree matching, wherein each image is matched against its nearest visual neighbors through a vocabulary tree. To build the vocabulary tree, we first compute the descriptor centroids using `KMeans++` [2], then `KDTree` [17] is used to build the vocabulary tree using VLAD [1] descriptors. The vocabulary tree serves as a visual database enabling retrieval of database images that closely resemble the query image in terms of visual appearance.

Enforcing panoramic constraints for bundle adjustment. In the conventional SfM workflow, unordered input image collections lead to the independent treatment of each image. However, in our scenario where perspective images stem from panorama sampling, we can leverage the known transformations or relative poses between frames that are sampled from the same panorama. To capitalize on this, we augment the typical Bundle Adjustment [52] (BA) by incorporating the known relative poses between frames within the same panorama. In our context, two perspective images from a common panorama are linked by a pure rotation around the z-axis (i.e., along the yaw direction, refer to Eq. 1). In particular, for a perspective image $\mathcal{I}_{ij}$ characterized by its extrinsic camera parameters $(R_{i,j}, t_{i,j})$, we introduce an additional optimization objective $\mathcal{L}_{\text{pano}} = \mathcal{L}_{\text{trans}} + \mathcal{L}_{\text{rot}}$, where:

$$\begin{aligned} \mathcal{L}_{\text{trans}} &= \sum_{i=1}^{N}\sum_{j=2}^{T} ||t_{i,j} - t_{i,j-1}||^2, \\ \mathcal{L}_{\text{rot}} &= \sum_{i=1}^{N}\sum_{j=2}^{T} ||R_{i,j}^{\top} R_{i,j-1} - R_z\left(\frac{2\pi}{T}\right)||^2 \\ \text{s.t.} \quad & R \in \text{SO}(3),\ t \in \mathbb{R}^3 \end{aligned} \tag{2}$$

where $R_z(\theta)$ denotes the rotation matrix around the z-axis by an angle of θ. As in COLMAP [47], Levenberg-Marquardt [23, 52] is used for optimization. As shown in Figure 3, this constraint helps correct erroneous camera poses during 3D reconstruction, particularly when working with a limited number of images.

Metric scale calibration and ground plane fitting. Using GPS coordinates of GSV panoramas, we geo-register the *up-to-scale* SfM reconstruction via a 3D similarity transformation optimized between the SfM coordinates and Earth-Centered-Earth-Fixed (ECEF) Cartesian coordinates. This results in a *metric scale* 3D scene reconstruction. Subsequently, the road plane is estimated by fitting a plane to the set of 3D points whose 2D pixel locations lie on the *road/lane markings* obtained from an off-the-shelf semantic segmentation method [11].

3.2. Camera Localization

Camera localization step aims to determine the intrinsic and extrinsic parameters of the traffic camera with respect to the 3D scene. As depicted in Figure 2, we adopt a visual localization pipeline that involves localizing the query image (from the traffic camera) within the 3D reconstruction constructed using GSV images in the previous step.

For every input query image, we retrieve the top-k similar database images, where k is a predefined value, from the vocabulary tree built in the 3D reconstruction step (Section 3.1). We then match the query image with the k retrieved database images to establish 2D-3D correspondences. For this, we use the learned feature matching method SuperGlue [45] with SuperPoint [13] feature descriptors to match the query image with the database images. Since our query image is uncalibrated, we follow a sampling-based approach [26, 47] where the pose/focal length is estimated using RANSAC and a minimal pose solver (e.g., [18, 30]). Finally, we perform an extra bundle adjustment step (with panoramic constraints as in Eq. 2) to refine both intrinsic parameters and camera poses.

Figure 4. Traffic camera image matched with GSV images using SuperPoint [13] and SuperGlue [45]. These methods provide reliable correspondences for accurate absolute pose estimation despite viewpoint and illumination differences.

It is worth noting that the use of learned feature matching is crucial in this step, as it has been shown to outperform hand-crafted feature descriptors and matching methods [45], particularly in cases where the viewpoint of the traffic camera (often much higher above ground) differs significantly from that of the Google Street View (GSV) images (captured from driving viewpoints). Utilizing SuperPoint [13] and SuperGlue [45] enables the generation of a large number of accurate matches between the query image and the comprehensive GSV database images (as shown in Figure 4). This rich set of matches allows robust recovery of both the intrinsic and extrinsic camera parameters. Recent advances in local feature matching, such as LightGlue [32], can potentially further improve efficiency.

4. Experimental Results

4.1. Calibrated Urban Traffic Cameras (CUTC) Dataset

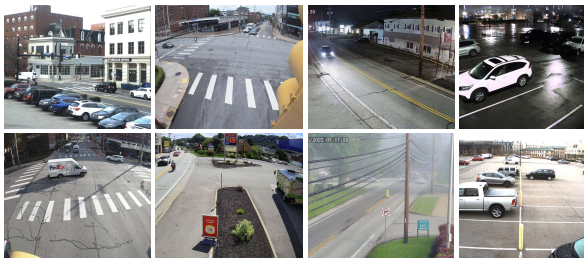


Figure 5. Example images of our Calibrated Urban Traffic Cameras (CUTC) dataset, with diverse scenes and viewpoints.

There are several existing datasets designed for evaluating traffic monitoring algorithms, notably BrnoCompSpeed [51] and Revaud et al. [43]. However, as mentioned in [43], BrnoCompSpeed [51] has limited diversity as the cameras mainly captures vehicles moving in straight lines on highways or freeways. In contrast, Revaud et al. [43] sought to address this limitation by introducing the CCTV dataset (denoted as Revaud-CCTV dataset), which better emulates real-life CCTV cameras' content and conditions such as low image resolution, non-straight roadways, and imperfect camera lenses. Although both BrnoCompSpeed and Revaud-CCTV provide ground-truth vehicle speeds, the lack of accurate camera calibration restricts its utility for novel traffic analysis applications.

To bridge this gap, we present a new dataset called *Calibrated Urban Traffic Cameras (CUTC)*. This dataset comprises 20 cameras distributed across 6 diverse locations within public urban settings. An example of our CUTC dataset is shown in Figure 5. Our team installed these cameras which underwent full calibration using checkerboard-based methods [56] (`OpenCV` calibration for `ChArUco` board) before deployment. Similar to the setup in [51], our dataset also incorporates manually measured markers (e.g., lane markings, crosswalks, etc.) on the road plane, with known dimensions between them. These ground-truth measurements serve as reference points for evaluating distance measurements on the ground plane (see Section 4.2).

4.2. Evaluation Metrics

Intrinsic Parameters: Considering the intrinsics parameters computed by checkerboard-based methods [56] as ground-truth, we report the mean error (in %) for focal lengths (f_x, f_y), principal points (p_x, p_y), and distortion coefficients (k_1, k_2, p_1, p_2) over 20 cameras in our CUTC dataset.

Ground Distance Measurements: Following [6, 28], using manually measured distances between pairs of points on

the road plane (e.g., lane markings, crosswalks, etc.) along with their pixel positions in the images, we then computed the normalized error in distance measurement, defined as $r_i = \frac{|\hat{d}_i - d_i|}{\hat{d}_i}$, where $\hat{d}_i$ is the i-th ground-truth distance measurement and d_i is the i-th measurement based on the ray-plane intersection using the estimated intrinsic matrix and ground-plane equation. This metric effectively gauges the accuracy of both intrinsic and extrinsic parameters.

4.3. Baseline Methods

We compare our method to SOTA automatic camera calibration approaches designed specifically for traffic cameras, including OptInOpt [3], PlaneCalib [4], DeepVPCalib [28], and Revaud et al. [43]. In particular, OptInOpt [3] and PlaneCalib [4] rely on localizing 2D landmarks with exact 3D CAD models to infer the focal length of the camera and vehicle poses, DeepVPCalib [28] relies on detecting pairs of vanishing points for multiple vehicles in a scene to obtain the focal length of the camera and the orientation of the road plane, and Revaud et al. [43] learns to predict the calibration (homography between image plane and ground plane) by training solely from synthetic 3D car models.

4.4. Quantitative Results

Method	f_x	f_y	p_x	p_y	k_1	k_2	p_1	p_2
OptInOpt [3]	11.72	11.21	2.57*	2.61*	✗	✗	✗	✗
PlaneCalib [4]	9.88	9.71	2.57*	2.61*	✗	✗	✗	✗
DeepVPCalib [28]	7.51	7.33	2.57*	2.61*	✗	✗	✗	✗
Ours	**3.17**	**3.54**	**2.11**	**2.02**	**7.56**	**8.28**	**6.71**	**8.43**

Table 1. Comparison between our approach and SOTA techniques in terms of mean error (in %) of focal lengths (f_x, f_y), principal points (p_x, p_y), and distortion coefficients (k_1, k_2, p_1, p_2). (✗: unavailable, *: method assumes principal point at image center.)

Method	Max Error (%)	Median Error (%)	RMSE (%)
OptInOpt [3]	15.78	10.80	12.87
PlaneCalib [4]	14.32	9.23	11.69
DeepVPCalib [28]	12.17	8.11	10.62
Revaud et al. [43]	14.87	10.91	12.54
Ours	**6.75**	**3.22**	**4.68**

Table 2. Comparison between our approach and SOTA automatic calibration techniques in terms of max, median, and RMSE (in %) between measured vs. estimated distances on ground plane.

Intrinsics Parameters: In Table 1, using checkerboard-based calibration as ground-truth, we compare our approach against SOTA automatic calibration techniques [3, 4, 28]. Our method significantly outperforms SOTA methods due to the fact that the dense coverage of GSV images (with *known* intrinsics parameters) and reliable correspondences from learned feature matching allows us to register the traffic camera into the 3D scene with high accuracy. Conversely, methods like OptInOpt [3] and PlaneCalib [4], relying on precise 3D CAD models of vehicles, exhibit diminished generalization accuracy when faced with cameras in diverse countries with unknown vehicle type. Additionally, we observe that the focal length estimation from DeepVPCalib [28] is instable as it is highly sensitive to the accuracy of detected vanishing points. Lastly, these methods assume no distortion, an assumption that may not hold in practice. Hence, an additional strength of our approach is its capability to estimate distortion parameters (radial and tangential) with a satisfactory accuracy level (within 10%).

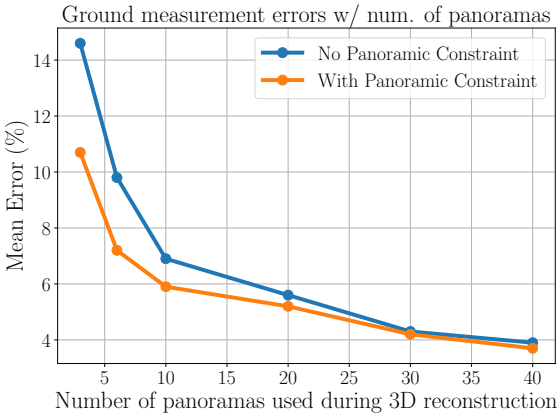


Figure 6. We show the mean errors on estimating ground distances w.r.t. the number of panoramas being used during 3D reconstruction. By enforcing panoramic constraint during reconstruction, our method improves the accuracy significantly over the baseline, especially in small number of views.

Distance Measurements: Using manually measured distances on the ground plane, in Table 2, we report the max, median, and root-mean-squared error (in %) over all of the possible pairs of ground truth measurements in our CUTC dataset. For methods that do not directly infer metric scene scale such as DeepVPCalib [28], we scale the estimated distances with the ground-truth scale. As shown in Table 2, our method outperforms existing SOTA methods by a large margin, demonstrating the accuracy of our camera calibration as well as estimated scene geometry. In our experiments, we observe limited generalizability from the pre-trained model of Revaud et al. [43] that was trained exclusively on synthetic 3D car models.

Enforcing the known relative pose between frames from the same panorama improves the accuracy of reconstruction: While it is well-known that using more images during 3D reconstruction leads to better camera calibration accuracy, our key observation is that by enforcing the known relative pose between frames from the same panorama during 3D reconstruction, we can significantly boost accuracy, especially when the number of GSV panoramas being used for reconstruction is limited (as

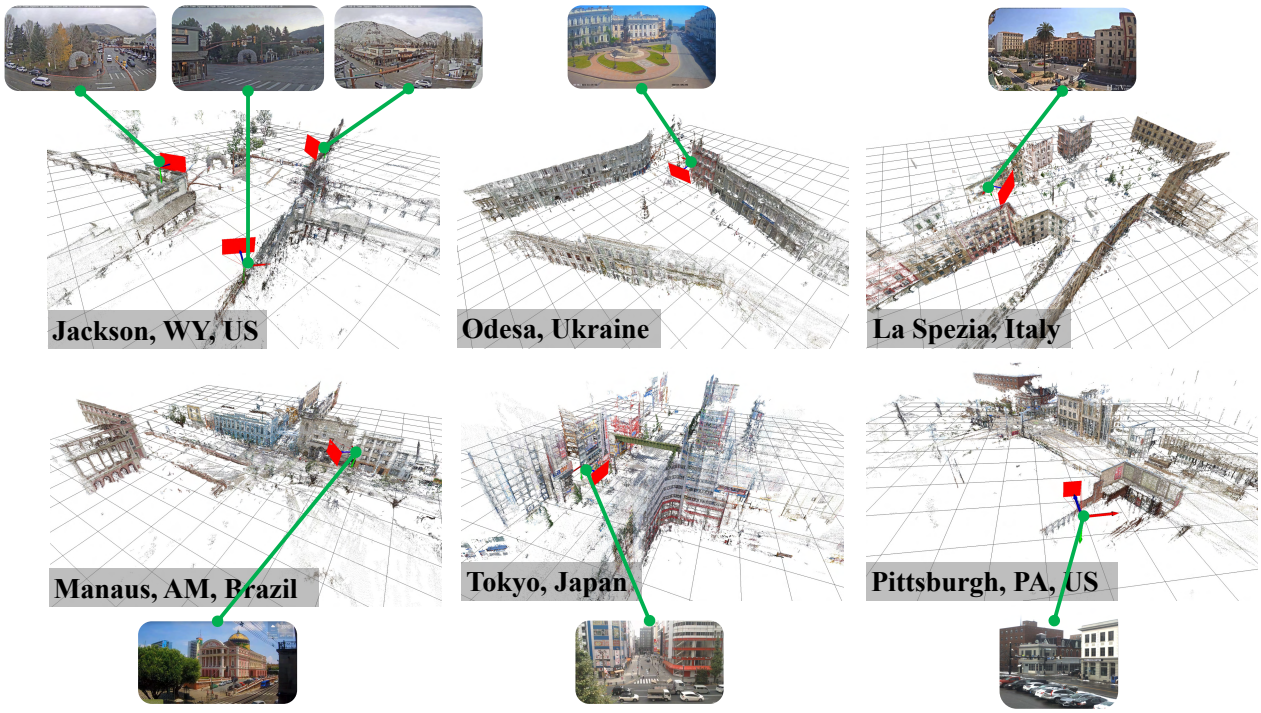


Figure 7. Additional examples demonstrate our method's robustness in reconstructing scenes and localizing cameras spanning various countries and continents (traffic camera visualized in red).

shown in Figure 6). This becomes crucial when dealing with the limited availability of GSV images at a location, or when computational efficiency is a priority.

4.5. Qualitative Results

Our camera localization method proves versatile across diverse cameras in real-world settings. As depicted in Figure 7, we successfully achieve both accurate 3D scene reconstruction and precise camera localization across different locations spanning multiple countries and continents. Importantly, our framework's adaptability extends beyond Google Street View, making it highly versatile for different street-level imagery sources [8, 38], with the potential to achieve camera calibration on a global scale.

5. Applications

With accurate camera calibration information, we show its applications in 3D reconstruction and speed measurements of moving vehicles, allowing us to gain unique insights that can inform decision-making processes related to traffic management and safety measures.

In Figure 8, we provide automatic vehicle speed estimates and activity heatmaps for two different cameras. First, we use an off-the-shelf object detector [25] and tracker [5] to compute the tracked detections of every vehicle. Using vehicles' active mean shape models and detected 2D keypoints [27, 42], following [31], we optimize for the 6DoF vehicle pose and shape variations (defined by PCA components of the mean shape model) for each vehicle track by enforcing all the detected objects to lie on the ground plane.

Activity Heatmap: Heatmaps visualize the level of vehicle activity at each camera location. These heatmaps are generated by aggregating the centroid of the 3D tracks of all vehicles over the entire data acquisition period, then normalized by the maximum count, resulting in a value ranging from 0 (blue, no activity) to 1 (red, high activity).

Vehicle Speed: For each camera, we created virtual speed traps (visualized as a green line on the road) that allows estimates of speed a vehicle crossing over the region of interest. Consequently, the reported speeds correspond to instantaneous speed readings garnered from the virtual speed trap. By leveraging the reconstructed 3D shape and pose of vehicles, we calculate the velocity as the front of the vehicle crosses the virtual speed trap on the ground plane. As depicted in Figure 8, precise camera calibration facilitates the accurate measurement of vehicle speeds, enabling us to automatically derive valuable insights from the data. For instance, the system can identify prevalent traffic patterns to enhance urban planning or detect anomalies such as accidents or high-speed emergency vehicles on duty. Importantly, the benefits of accurate metric 3D scene reconstruction and camera calibration extend beyond speed estimation as they are crucial for various applications, includ-

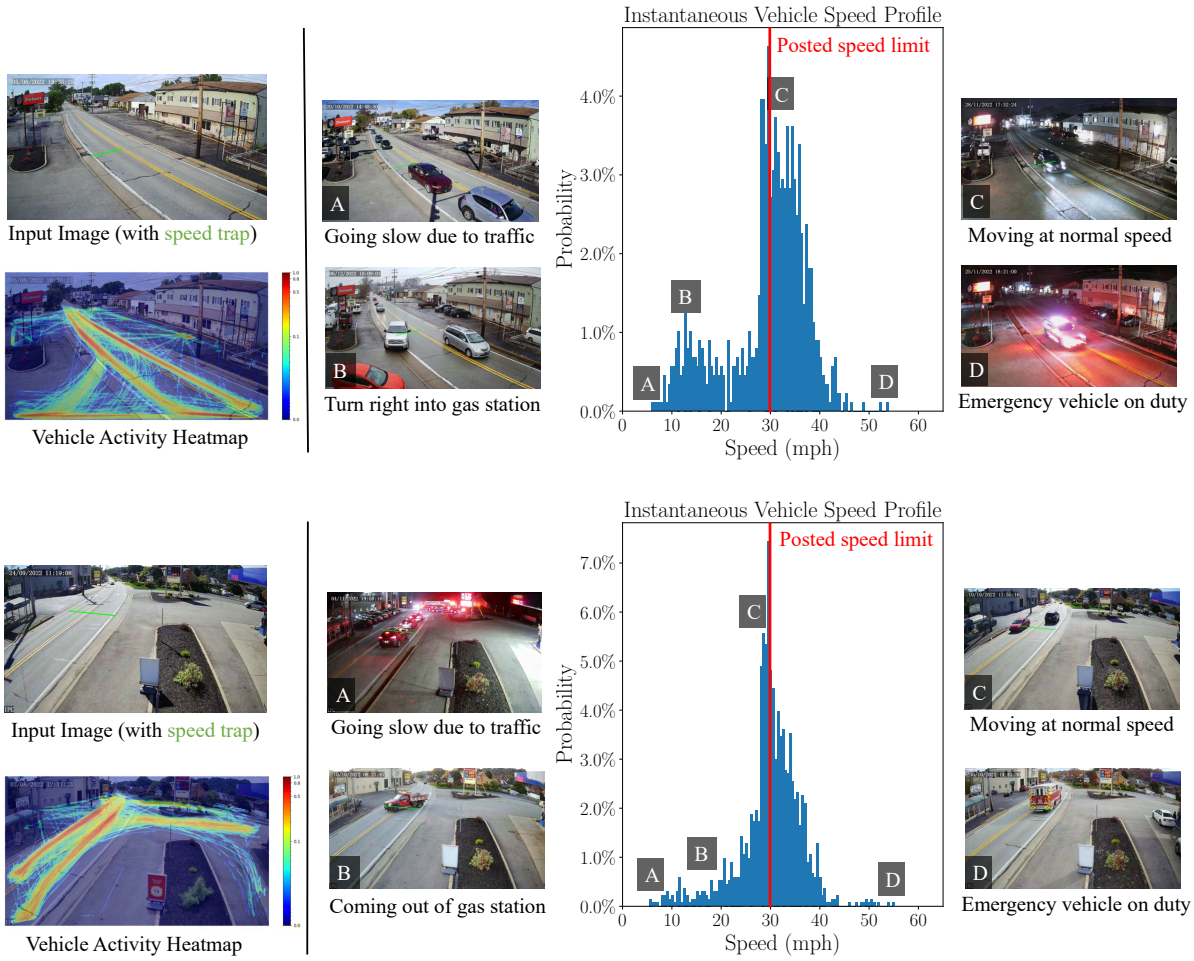


Figure 8. Speed estimates and activity heatmaps depicted for two distinct virtual speed traps (indicated by green lines) across two different cameras. Various scenarios are displayed, including slow traffic due to congestion and fast movement of emergency vehicles.

ing understanding human-vehicle interaction for accident prediction and prevention [33], achieving multi-camera fusion by aligning different cameras' views to a common frame [6, 39], and so on.

6. Discussion

We have presented a scalable framework that leverages street-level imagery for metric 3D model reconstruction, enabling accurate calibration of real-world traffic cameras. Our approach can be applied to any camera with sufficient nearby street-level imagery, making it practical to be used worldwide. We show the framework's value in traffic analysis through insights derived from 3D vehicle reconstruction and speed measurement, providing valuable information for improving transportation systems and urban infrastructure.

Potential Societal Impact: We do not perform any human subject studies from these cameras. To preserve the privacy of the object captured in the images, we blur the faces and license plates in all the images to be released. This study is designated as non-human subjects research by our Institutional Review Board (IRB).

Limitations: Our method requires capturing at least some portion of the scene's "background" for feature matching. Thus, scenes with severely limited contextual information, e.g., situations where cameras are oriented to solely capture freeway surfaces while looking straight down, can hinder the performance of our approach.

Acknowledgements: This work was supported in part by an NSF Grant CNS-2038612, a DOT RITA Mobility-21 Grant 69A3551747111, and Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *CVPR*, 2013. 4

[2] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *SODA*, 2007. 4

[3] Vojtěch Bartl and Adam Herout. Optinopt: Dual optimization for automatic camera calibration by multi-target observations. In *AVSS*, 2019. 6

[4] Vojtěch Bartl, Roman Juranek, Jakub Špaňhel, and Adam Herout. Planecalib: Automatic camera calibration by multiple observations of rigid objects on plane. In *DICTA*, 2020. 2, 3, 6

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 7

[6] Romil Bhardwaj, Gopi Krishna Tummala, Ganesan Ramalingam, Ramachandran Ramjee, and Prasun Sinha. Autocalib: Automatic traffic camera calibration at scale. *TOSN*, 2018. 2, 5, 8

[7] Huikun Bi, Zhong Fang, Tianlu Mao, Zhaoqi Wang, and Zhigang Deng. Joint prediction for kinematic trajectories in vehicle-pedestrian-mixed scenes. In *CVPR*, 2019. 1

[8] Bing. Bing Streetside. https://www.bing.com/maps/. 2, 7

[9] Oleksandr Bogdan, Viktor Eckstein, Francois Rameau, and Jean-Charles Bazin. Deepcalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–10, 2018. 2

[10] Joseph R Cathey and Matthew A Dailey. Camera calibration using lane markings: An evaluation of vanishing point detection methods. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):124–133, 2005. 1, 2

[11] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 4

[12] Matthew A Dailey, Benjamin C Schoepflin, Juraj Sochor, and Michal Seman. Camera calibration for traffic scene analysis using vehicle motion. *IEEE Transactions on Intelligent Transportation Systems*, 1(1):43–50, 2000. 2

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2, 4, 5

[14] Hoang Dung Do and Reinhard Klette. Camera calibration for road scene analysis using vehicle motion and lane markings. *IEEE Transactions on Intelligent Transportation Systems*, 16(7):2700–2712, 2015. 2

[15] Katerina Dubska, Jiri Matas, Ondrej Holik, and Michal Seman. Camera calibration using vehicle motion. *IEEE Transactions on Intelligent Transportation Systems*, 15(1):283–294, 2014. 2

[16] Katerina Dubska, Jiri Matas, Ondrej Holik, and Michal Seman. Camera calibration for road scene analysis using vehicle motion with robust estimation of camera parameters. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):540–551, 2015. 2

[17] Jerome H Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):209–226, 1977. 4

[18] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 25(8):930–943, 2003. 5

[19] Panagiotis Giannakeris, Vagia Kaltsa, Konstantinos Avgerinakis, Alexia Briassouli, Stefanos Vrochidis, and Ioannis Kompatsiaris. Speed estimation and abnormality detection from surveillance cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 93–99, 2018. 1

[20] Google. Celebrating 15 years of Street View. https://www.google.com/streetview/anniversary/. 3

[21] Google. Google Street View. https://www.google.com/streetview/. 2, 3

[22] Vasileios Grammatikopoulos, Manolis N Lourakis, and John K Tsotsos. Camera calibration for road scene analysis using vanishing points. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):134–144, 2005. 2

[23] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[24] Jiebo He and Nicholas Yung. Camera calibration for traffic scene analysis using lane markings. *IEEE Transactions on Intelligent Transportation Systems*, 8(3):417–427, 2007. 2

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 7

[26] Arnold Irschara, Christopher Zach, Jan-Michael Frahm, and Horst Bischof. From structure-from-motion point clouds to fast location recognition. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2599–2606. IEEE, 2009. 5

[27] Lei Ke, Shichao Li, Yanan Sun, Yu-Wing Tai, and Chi-Keung Tang. Gsnet: Joint vehicle pose and shape reconstruction with geometrical and scene-aware supervision. In *European Conference on Computer Vision*, pages 515–532. Springer, 2020. 7

[28] Viktor Kocur and Milan Ftáčnik. Traffic camera calibration via vehicle vanishing point detection. In *ICANN*, pages 628–639. Springer, 2021. 2, 3, 5, 6

[29] Man Lan, Jiang Zhao, and Tiantian Zhu. Camera calibration for traffic scene analysis using multiple vanishing points. *IEEE Transactions on Intelligent Transportation Systems*, 15(4):1806–1817, 2014. 2

[30] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009. 5

[31] Fangyu Li, N Dinesh Reddy, Xudong Chen, and Srinivasa G Narasimhan. Traffic4d: Single view longitudinal 4d reconstruction of repetitious activity using self-supervised experts. In *IEEE Intelligent Vehicles Symposium*, 2021. 7

[32] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 5

[33] Franz Loewenherz, Victor Bahl, and Yinhai Wang. Video analytics towards vision zero. *Institute of Transportation Engineers. ITE Journal*, 87(3):25, 2017. 8

[34] Manuel Lopez, Roger Mari, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11817–11825, 2019. 2

[35] Luiz Henrique Luvizon, Marcelo de Souza, and Mohamed Bennamoun. Camera calibration for traffic scene analysis using vehicle motion. *IEEE Transactions on Intelligent Transportation Systems*, 15(1):363–374, 2014. 2

[36] Luiz Henrique Luvizon, Marcelo de Souza, and Mohamed Bennamoun. Camera calibration for traffic scene analysis using vehicle motion with uncertainty estimation. *IEEE Transactions on Intelligent Transportation Systems*, 17(1):270–281, 2016. 2

[37] Ricardo Maduro and Reinhard Klette. Camera calibration for road scene analysis using manual measurements. *IEEE Transactions on Intelligent Transportation Systems*, 9(3):501–510, 2008. 2

[38] Mapillary. Mapillary Maps. https://www.mapillary.com/. 2, 7

[39] K. Muller, A. Smolic, M. Drose, P. Voigt, and T. Wiegand. 3-d reconstruction of a dynamic environment with a fully calibrated background for traffic scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(4):538–549, 2005. 8

[40] Julian Nubert, Nicholas Giai Truong, Abel Lim, Herbert Ilhan Tanujaya, Leah Lim, and Mai Anh Vu. Traffic density estimation using a convolutional neural network. *arXiv preprint arXiv:1809.01564*, 2018. 1

[41] Angga Nurhadiyatna and Reinhard Klette. Camera calibration for road scene analysis using manual measurements. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1149–1159, 2013. 2

[42] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7326–7335, 2019. 7

[43] Jerome Revaud and Martin Humenberger. Robust automatic monocular vehicle speed estimation for traffic surveillance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4551–4561, 2021. 3, 5, 6

[44] Imad Sabbani, Andres Perez-Uribe, Omar Bouattane, and Abdellah El Moudni. Deep convolutional neural network architecture for urban traffic flow estimation. *IJCSNS International Journal of Computer Science and Network Security*, 2018. 1

[45] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 2, 4, 5

[46] Benjamin C Schoepflin and Matthew A Dailey. Camera calibration for traffic scene analysis using vehicle motion. *IEEE Transactions on Intelligent Transportation Systems*, 4(2):111–120, 2003. 2

[47] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 4, 5

[48] Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–9. IEEE, 2018. 1

[49] Mostafa Sina and Reinhard Klette. Camera calibration for road scene analysis using manual measurements. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1595–1604, 2013. 2

[50] Jakub Sochor, Roman Juránek, and Adam Herout. Traffic surveillance camera calibration by 3d model bounding box alignment for accurate vehicle speed measurement. *Computer Vision and Image Understanding*, 161:87–98, 2017. 2

[51] Jakub Sochor, Roman Juránek, Jakub Špaňhel, Lukáš Maršík, Adam Širokỳ, Adam Herout, and Pavel Zemčík. Comprehensive data set for automatic single camera visual speed measurement. *IEEE Transactions on Intelligent Transportation Systems*, 20(5):1633–1643, 2018. 1, 2, 5

[52] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings*, pages 298–372. Springer, 2000. 4

[53] Roger Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal on Robotics and Automation*, 3(4):323–344, 1987. 1, 2

[54] Kunfeng Wang, Hua Huang, Yuantao Li, and Fei-Yue Wang. Research on lane-marking line based camera calibration. In *2007 IEEE International Conference on Vehicular Electronics and Safety*, pages 1–6. IEEE, 2007. 1, 2

[55] Xingchen Zhang, Yuxiang Feng, Panagiotis Angeloudis, and Yiannis Demiris. Monocular visual traffic surveillance: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14148–14165, 2022. 3

[56] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000. 1, 2, 5