CVLNet: Cross-View Semantic Correspondence Learning for Video-based Camera Localization

Yujiao Shi[1], Xin Yu[2], Shan Wang[1], Hongdong Li[1]

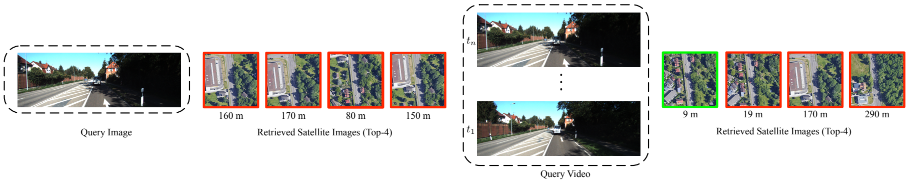
[1]Australian National University [2]University of Technology Sydney

Abstract. This paper tackles the problem of Cross-view Video-based camera Localization (CVL). The task is to localize a query camera by leveraging information from its past observations, *i.e.*, a continuous sequence of images observed at previous time stamps, and matching them to a large overhead-view satellite image. The critical challenge of this task is to learn a powerful global feature descriptor for the sequential ground-view images while considering its domain alignment with reference satellite images. For this purpose, we introduce CVLNet, which first projects the sequential ground-view images into an overhead view by exploring the ground-and-overhead geometric correspondences and then leverages the photo consistency among the projected images to form a global representation. In this way, the cross-view domain differences are bridged. Since the reference satellite images are usually pre-cropped and regularly sampled, there is always a misalignment between the query camera location and its matching satellite image center. Motivated by this, we propose estimating the query camera's relative displacement to a satellite image before similarity matching. In this displacement estimation process, we also consider the uncertainty of the camera location. For example, a camera is unlikely to be on top of trees. To evaluate the performance of the proposed method, we collect satellite images from Google Map for the KITTI dataset and construct a new cross-view video-based localization benchmark dataset, KITTI-CVL. Extensive experiments have demonstrated the effectiveness of video-based localization over single image-based localization and the superiority of each proposed module over other alternatives.

1 Introduction

Cross-view image-based localization using ground-to-satellite image matching has attracted significant attention these days [1–11]. It has found many practical applications such as autonomous driving and robot navigation. Prior works have been focused on localizing omnidirectional ground-view images with a 360° Field-of-View (FoV), which helps to provide rich and discriminative features for localization. However, when a regular forward-looking camera with a limited FoV is used, those omnidirectional camera-based algorithms suffer severe performance degradation.

To tackle this challenge, this paper proposes to use a continuous short video, *i.e.*, a sequence of ground-view images, as input for the task of visual localization.


Query Image
160 m
170 m
80 m
150 m
Retrieved Satellite Images (Top-4)
t_n
t_1
Query Video
9 m
19 m
170 m
290 m
Retrieved Satellite Images (Top-4)

(a) Cross-view image-based localization (b) Cross-view video-based localization

Fig. 1: Single-frame image-based localization (a) Vs. Multi-frame video-based localization (b). The multi-frame video-based localization leverages richer scene context of a query place, increasing the discriminating power of query descriptors compared to single image-based localization. As a result, the matching satellite image, marked by green border, from the database is more likely to be retrieved. Red border indicates non-matching satellite images to the query image.

Specifically, we localize a camera at the current time stamp t_n by augmenting it with previous observations at time *i.e.*, $t_1 \sim t_{n-1}$, as shown in Fig. 1. Compared to using a single query image, a short video provides richer visual and dynamic information about the current location.

We present a Cross-view Video-based Localization Network, named CVLNet, to address the camera localization problem. To the best of our knowledge, our CVLNet is the first vision- and deep-based cross-view geo-localization framework that exploits a continuous video rather than a single image to pinpoint the camera location.

Our CVLNet is composed of two branches that extract deep features from ground and satellite images, respectively. Considering the drastic viewpoint changes between the two-view images, we first introduce a Geometry-driven View Projection (GVP) module to transform ground-view features to the overhead view by explicitly exploring their geometric correspondences. Then, we design a Photo-consistency Constrained Sequence Fusion (PCSF) module to fuse the sequential features. Our PCSF first estimates the reliability of the sequential ground-view features in overhead view by leveraging photo-consistency across them and then aggregates them as a global query descriptor. In this manner, we achieve more discriminative and reliable ground-view feature representation.

Since satellite images in a database are usually pre-cropped and sampled at discretized locations, there would be a misalignment between a query camera location and its matching satellite image center. Furthermore, a query camera is usually impossible in some regions (*e.g.*, on top of a tree), while likely on the other areas (*e.g.*, road). Hence, we propose a Scene-prior driven Similarity Matching (SSM) strategy to estimate the relative displacement between a query camera location and a satellite image center while restricting the search space by scene priors. The scene priors are learned statistically from training rather than pre-defined. With the help of SSM, our CVLNet can eliminate unreasonable localization results.

In order to train and evaluate our method, we curate a new cross-view dataset by collecting satellite images for the KITTI dataset [12] from Google Map [13]. The new dataset combines sequential ground-view images from the original KITTI dataset and the newly collected satellite images. To the best of our knowledge, it is not only the first cross-view video-based localization dataset, but also the first cross-view localization dataset where ground-view images are captured by a perspective pin-hole camera with a restricted FoV (rather than being cropped from Google street-view panoramas [1, 8]). Extensive experiments on the newly collected dataset demonstrate that our method effectively localizes camera positions and outperforms the state-of-the-art remarkably.

2 Related Work

Image-based localization. The image-based localization problem is initially tackled as a ground-to-ground image matching [14–19], where both the query and database images are captured at the ground level. However, those methods cannot localize query images when there is no corresponding reference image in the database. Thanks to the wide-spread coverage and easy accessibility of satellite imagery, recent works [20–22, 1, 23, 2–6, 24, 7–11, 25–29] resort to satellite images for city-scale localization.

While recent works on city-scale ground-to-satellite localization have achieved promising results, they mostly focus on localizing isolated omnidirectional ground images. When the query camera has a limited FoV, we propose using a continuous video instead of a single image for camera localization, improving the discriminativeness of the query location representation.

Video-based localization. The concept of video-based localization can be divided into three main categories; Visual Odometry (VO) [30–32], Visual-SLAM (vSLAM) [33–38] and Visual Localization [39–45]. VO techniques can be classified according to their camera setup — either monocular or stereoscopic or their processing techniques — either feature-based or appearance-based. VO methods usually use a combination of feature tracking and feature matching [46, 47]. vSLAM pertains to simultaneously creating a map of features and localizing the robot in that map, all using visual information [48, 49]. Many a time, the map is pre-built, and the robot needs to localize itself using camera-based map-matching, which is referred to as Visual Localization [50]. Even though these methods use a series of image frames to determine the robot's location, they match information from the same viewpoint. In our work, we have developed a cross-view video-based localization approach by leveraging a sequence of images with varied viewpoints and limited FoVs, aiming to improve the representativeness of a query location significantly.

Multi-view Counting/Detection. There are also related methods which project images to the same ground-plane for fusion, such as multi-view counting [51–54], multi-view detection [55–59] methods. We share the similar idea of projecting features on the ground plane, but solve different downstream tasks and have distinct challenges.

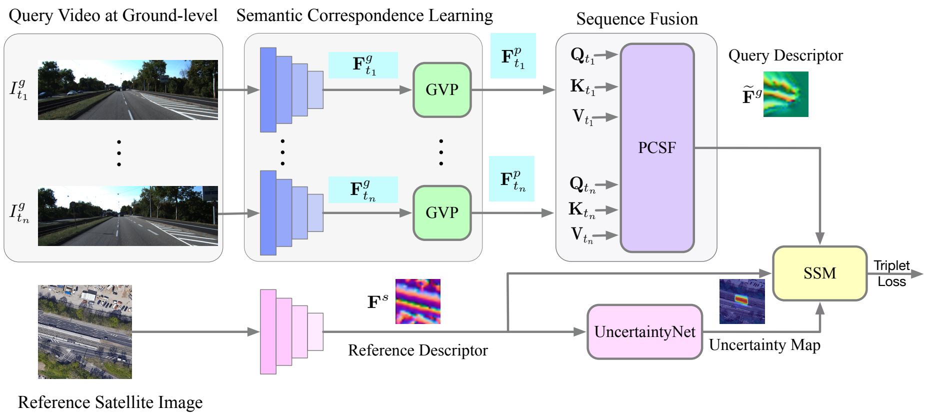


Fig. 2: Overview of our proposed CVLNet. Our Geometry-driven View Projection (GVP) module first aligns the sequential ground-view features in the overhead view and presents them in a unified coordinate system. Next, the Photo-consistency Constrained Sequential Fusion (PCSF) module measures the photo-consistency of an overhead view pixel across the different ground-views and fuses them together, obtaining a global feature representation $\widetilde{\mathbf{F}}^g$ of the query video. The global feature representation is then compared with the satellite feature map $\mathbf{F}^s$ with a Scene-prior driven Similarity Matching (SSM) scheme to determine the relative displacement between the query camera location and the satellite image center, guided by an uncertainty map. After alignment, the feature similarity is then computed for image retrieval.

3 CVLNet: Cross-view Video-based Localization

This paper tackles the ground-to-satellite localization task. Instead of using a single query image captured at the ground level, we augment the query image with a short video containing previous observations. To solve this task, our motivation is first projecting the images in the ground video to an overhead[1] perspective and then extracting a global description from the projected image sequence for localization. An overview of our pipeline is illustrated in Fig. 2.

3.1 Geometry-driven view projection (GVP)

Prior methods often resort to a satellite to ground projection to bridge the cross-view domain gap. This is achieved either by a polar transform [6, 8, 10] or a projective transform [60, 25, 61]. However, both transforms need to know the query camera location with respect to the satellite image center. In the CVUSA and CVACT dataset where polar transform performs excellent, the query images accidentally align with their matching satellite image center, which however does not occur in practice. When there is a large offset between the real camera location and its assumed location with respect to its matching satellite image (*e.g.*, satellite image center in polar transform), the performance will be impeded significntly. Hence, instead of projecting satellite images to ground views, we introduce a Geometry-driven View Projection (GVP) module to transform ground-view images to overhead view.

[1] For clarity, we use "overhead" throughout the paper to denote the projected features from ground-views, and "satellite" to indicate the real satellite image/features.

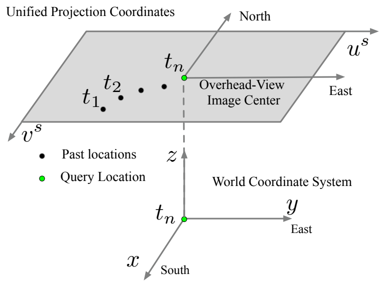


Fig. 3: A unified projection coordinates for the projected overhead-view features. Ground-view observations at timestamps $T = \{t_1, t_2, \ldots, t_n\}$ are projected to the same overhead-view grid with the center corresponding to the query camera location at t_n.

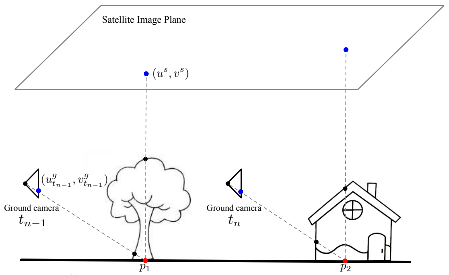


Fig. 4: Geometry-driven cross-view semantic correspondence learning. Both tree canopy and tree trunk are "trees". Building roof and facades are both "buildings".

Starting from a blank canvas in the overhead view with its center corresponds to the geospatial location of the query camera, we aim to fill it with features collected from ground-view images. We set the origin of the world coordinate system to the geo-spatial query camera location as well, with its x-axis pointing to the south direction, y axis pointing to the east direction, and the z-axis vertically upward. Different ground-view images in a video sequence are projected to the same overhead-view coordinate system so that they are geographically aligned after projection. Fig. 3 provides a visual illustration of the coordinate systems.

Parallel projection of a satellite camera. The projection between the satellite image coordinate system (u^s, v^s) and the world coordinate system (x, y, z) can be approximated as a parallel projection [60], $[x, y]^T = \lambda[v^s - v^s_0, u^s - u^s_0]^T$, where (u^s_0, v^s_0) indicates the satellite map center, λ indicates the real-world distance between two neighboring pixels in the satellite map.

Perspective projection of ground-view images. Denote $\mathbf{R}_{t_i}$ and $\mathbf{t}_{t_i}$ as the rotation and translation for the camera at time step t_i in the world coordinate, $\mathbf{E}_{t_i}$ as the camera intrinsic, and N as the sequence number. The relative $\mathbf{R}_{t_i}$ and $\mathbf{t}_{t_i}$ can be easily obtained by Structure from Motion [62]. The projection between the world coordinate system (x, y, z) and the ground-view camera coordinate system $(u^g_{t_i}, v^g_{t_i})$ is expressed as $w_{t_i}[u^g_{t_i}, v^g_{t_i}, 1]^T = \mathbf{E}_{t_i}[\mathbf{R}_{t_i}, \mathbf{t}_{t_i}][x, y, z, 1]^T$, where w_{t_i} is a scale factor in the perspective projection.

Ground-to-satellite projection. There is a height ambiguity of satellite pixels in the ground-to-satellite projection. Instead of explicitly estimating the heights, we present a simple yet effective solution. Specifically, we project ground-view observations to the overhead view assuming satellite pixels lie on the ground plane. Rather than projecting original image RGB pixels, we project high-level deep features. The geometric projection from the ground-view to the overhead-view is derived as,

$$w_{t_i}[u^g_{t_i}, v^g_{t_i}, 1]^T = \mathbf{E}_{t_i}[\mathbf{R}_{t_i}, \mathbf{t}_{t_i}][\lambda(v^s - v^s_0), \lambda(u^s - u^s_0), -h, 1]^T. \tag{1}$$

where h is the height of the query camera with respect to the ground plane, and w_{t_i} can be computed from the above equation.

Denote $\mathbf{F}^g_{t_i} \in \mathbb{R}^{H \times W \times C}$ as ground-view image features by a CNN backbone, where H, W and C are the height, width and channels of the features, respectively, and $\mathrm{GVP}(\cdot)$ as the geometry-driven view projection operation illustrated in Eq. (1). The projected features in overhead view are then obtained by $\mathbf{F}^p_{t_i} = \mathrm{GVP}(\mathbf{F}^g_{t_i})$, $\mathbf{F}^p_{t_i} \in \mathbb{R}^{S \times S \times C}$, where S indicates the overhead-view feature map resolution.

This projection establishes the exact geometric correspondences between the ground and overhead views for scene contents on the ground plane. For scene objects with higher heights, projecting features rather than image pixels can alleviate the strict constraint while providing a cue that corresponding objects exist between the views. As shown in Fig. 4, for pixel (u^s, v^s), the projected feature from the ground-view at t_{n-1} represents the tree trunk, but the feature in the satellite image corresponds to the tree canopy. Both tree canopy and tree trunk indicate there is a tree at location p_2. Then, by applying a matching loss between the two features (tree trunk and tree canopy), the network will be trained to learn viewpoint invariant features, *i.e.*, both tree trunk and tree canopy are mapped to the semantic features of "tree".

The coverage of the canvas for ground-to-satellite projection is set to the reference satellite image coverage, *i.e.*, around 100m × 100m, with its center corresponding to the query camera location. When the sequence is too long with some previous image contents exceeding the canvas's pre-set coverage, the exceeded contents will not be collected. This is because scene contents that are too far from the query camera location are less important for localization, and it is better to cover most of the synthetic overhead-view feature map by referencing satellite images.

3.2 Photo-consistency constrained sequence fusion

We leverage photo consistency among different ground-views for the video sequence fusion. For a satellite pixel, when its corresponding features in several (more than two) ground views are similar, the existence of a scene object at this geographical location is highly reliable for these ground-views. We should highlight these corresponding features when generating descriptors for scene contents. Driven by this, we design a Photo-consistency Constrained Sequence Fusion (PCSF) module. Our PCSF module employs an attention mechanism [63] to emphasize reliable features in fusing a video sequence and obtaining a global descriptor for the video.

Our GVP block has aligned the original ground-view features at different time steps in a unified overhead-view coordinate. When the features of a geographical location observed by different ground views are similar, those features should be more reliable for localization. We leverage the self-attention mechanism [63] to measure the photo-consistency/similarity across different views and find reliable features. Specifically, for each projected feature map $\mathbf{F}^p_{t_i}$ at time step t_i, we compute its query, key and value by two stacked convolutional layers, denoted by

$\mathbf{Q}_{t_i}, \mathbf{K}_{t_i}, \mathbf{V}_{t_i} \in \mathbb{R}^{S\times S\times C}$, respectively. The stacked convolutional layers increase the receptive field and the representative ability of the key, query, and value features at each spatial location. Next, we compute the similarities between each projected feature map at t_i and other projected feature maps at t_j, $i, j = 1, ..., N$, and normalize them across all possible j by a softmax operation, expressed as,

$$\mathbf{M}_{i,j} = \mathrm{Softmax}_j \left(\mathbf{Q}_{t_i}^T \mathbf{K}_{t_j}\right), \quad \mathbf{M} \in \mathbb{R}^{N\times N\times S\times S}. \tag{2}$$

The final fused feature is obtained by,

$$\widetilde{\mathbf{F}}^g = \frac{1}{N}\sum\nolimits_i^N \sum\nolimits_j^N \mathbf{M}_{i,j}\mathbf{V}_{t_j}, \quad \widetilde{\mathbf{F}}^g \in \mathbb{R}^{S\times S\times C}. \tag{3}$$

In this way, we highlight the common features across the views and make the global descriptor reliable.

3.3 Scene-prior driven similarity matching

We want to address the location misalignment between a query camera location and its matching satellite image center by ground-to-satellite projection and spatial correlation between the projected features and the real satellite features. Hence, the satellite feature descriptors should be translational equivariant, which is an inherent property of conventional CNNs. Following most previous works [2, 6–8, 11], we use VGG16 [64] as our backbone for satellite (and ground) feature extraction. The extracted satellite features, denoted as $\mathbf{F}^s \in \mathbb{R}^{S\times S\times C}$, share the same spatial scale as the global representation of the query video. Next, we adopt a Normalized spatial Cross-Correlation (NCC) to estimate latent alignment between the query location and a satellite image center.

Denote $[\mathbf{F}^s]_{m,n}$ as a shifted version of a satellite feature map with its center at (m, n) in the original satellite feature map, and $m = 0, n = 0$ correspond to the center of the original satellite feature map. The similarity between $\mathbf{F}^s$ and $\widetilde{\mathbf{F}}^g$ aligned at (m, n) computed by NCC is,

$$\mathbf{D}_0(\mathbf{F}^s, \widetilde{\mathbf{F}}^g)_{m,n} = \frac{[\mathbf{F}^s]_{m,n} \cdot \widetilde{\mathbf{F}}^g}{\|[\mathbf{F}^s]_{m,n}\|_2 \|\widetilde{\mathbf{F}}^g\|_2}, \tag{4}$$

where $\mathbf{D}_0(\mathbf{F}^s, \widetilde{\mathbf{F}}^g) \in \mathbb{R}^{h\times w}$ denotes the similarity matrix between $\mathbf{F}^s$ and $\widetilde{\mathbf{F}}^g$ at all possible spatial-aligned locations, $m \in [-\frac{h}{2}, \frac{h}{2}]$, and $n \in [-\frac{w}{2}, \frac{w}{2}]$. A potential spatial-aligned location of the satellite map lies in a region of 10×10 m^2 in our KITTI-CVL dataset, as the database satellite image is collected very ten meters.

To exclude impossible query camera locations, *e.g.*, top of trees, we estimate an uncertainty map from the satellite semantic features, $\mathbf{U}(\mathbf{F}^s) = \mathcal{U}(\mathbf{F}^s)$, $\mathbf{U}(\mathbf{F}^s) \in \mathbb{R}^{h\times w}$, where $\mathcal{U}(\cdot)$ is the uncertainty net, composed of a set of convolutional layers. The value of each element in $\mathbf{U}(\mathbf{F}^s)$ is within the range of $[0, 1]$, forced by a Sigmoid layer. By encoding the uncertainty, The similarity between $\mathbf{F}^s$ and $\widetilde{\mathbf{F}}^g$ aligned at (m, n) is then written as,

$$\mathbf{D}(\mathbf{F}^s, \widetilde{\mathbf{F}}^g)_{m,n} = \frac{\mathbf{D}_0(\mathbf{F}^s, \widetilde{\mathbf{F}}^g)_{m,n}}{\mathbf{U}(\mathbf{F}^s)_{m,n}}. \tag{5}$$

When the uncertainty at (m, n) is large, the similarity between $\mathbf{F}^s$ and $\widetilde{\mathbf{F}}^g$ aligned at this location will be decreased. We do not have explicit supervisions for the uncertainty map. Rather, it is learned statistically from training. The relative displacement between $\mathbf{F}^s$ and $\widetilde{\mathbf{F}}^g$ is obtained by,

$$m^*, n^* = \arg\max_{m,n} \mathbf{D}(\mathbf{F}^s, \widetilde{\mathbf{F}}^g)_{m,n}. \tag{6}$$

During inference, we have no idea which one is the matching reference image for a query image. Thus the uncertainty-guided similarity matching is applied to all reference features (including non-matching ones). Furthermore, it is more challenging when a similarity score between non-matching ground and satellite features is high. Hence, we apply the similarity matching scheme to the pairs of query and non-matching reference images as well during training and minimize their maximum similarity, making the learned features more discriminative.

3.4 Training objective

We employ the soft-weighted triplet loss [2] to train our network. The loss includes a positive term to maximize the similarity between the matching query and reference pairs and a negative term to minimize the similarity between non-matching pairs. The non-matching term also prevents our view projection module from trivial solutions. Therefore, it is formulated as,

$$\mathcal{L} = \log\left(1 + e^{\alpha\left(d(\widetilde{\mathbf{F}}^g, \mathbf{F}^s) - d(\widetilde{\mathbf{F}}^g, \mathbf{F}^{s^*})\right)}\right), \tag{7}$$

where $\mathbf{F}^s$ is the matching satellite image feature to the ground feature $\mathbf{F}^g$, $\mathbf{F}^{s^*}$ is the non-matching satellite image feature, $d(\cdot, \cdot)$ is the L_2 distance between its two inputs after alignment, and α is set to 10.

4 The KITTI-CVL Dataset

KITTI is one of the widely used benchmark datasets for testing computer vision algorithms for autonomous driving [12]. In this paper, we intend to investigate a method for using a short video sequence for satellite image-based camera localization. For this purpose, we supplement the KITTI drive sequences with corresponding satellite images. This is done by cropping high-definition Google earth satellite images using the KITTI-provided GPS tags for vehicle trajectories. Based on these GPS tags of the ground-view images, we select a large region that covers the vehicle trajectory. We then uniformly partition the region into overlapping satellite image patches. Each satellite image patch has a resolution of 1280×1280 pixels, amounting to about 20 cm per pixel.

Training, Validation and Test sets. The KITTI data contains different trajectories captured at different time. In our Training, Validation and Test set split, the images of Training and Validation set are from the same region. The

Table 1: Query image numbers in the Training, Validation and Tests sets.

	Training	Validation	Test-1	Test-2
Distractor	✗	✗	✗	✓
Query Num	23,905	2,362	2,473	2,473

Validation set is constructed in this way to select the best model during training. In contrast, the images in the test set are captured at different regions from the Training and Validation sets. The test set aims to evaluate the generalization ability of the compared algorithms.

Only the nearest satellite image for each ground image in the sampled grids is retained for the Training and Validation set. We use the same method to construct our first test set, Test-1. Furthermore, we construct the second test set, Test-2, where all satellite images in the sampled grids are reserved. In other words, Test-2 contains many distracting satellite images, and it considers the real deployment scenario compared to Test-1. Visual illustrations of the differences between Test-1 and Test-2 are provided in the supplementary material. Tab. 1 presents the query ground image numbers of the Training, Validation, Test-1, and Test-2 sets.

5 Experiments

Evaluation metrics. Following the previous cross-view localization work [3], we use the distance and recall at top k ($r@k$) for the performance evaluation. Specifically, when one of the retrieved top k reference images is within 10 meters to the query ground location, it is regarded as a successful localization. The percentage of successfully localized query images is recorded as recall at top k. we set k to 1, 5, 10 and 100, respectively.

Implementation details. The input satellite image size is 512×512, center cropped from the collected images. The coverage of them is approximately $102\text{m} \times 102\text{m}$. The ground image resolution is 256×1024. The sizes of our global descriptor for query videos and satellite images are both 4096, which is a typical descriptor dimension in image retrieval. We follow prior arts [2–11] to adopt an exhaustive mini-batch strategy [1] with a batch size of $B = 8$ to prepare the training triplets. The Adam optimizer [65] with a learning rate of 10^{-4} is employed, and our network is trained end-to-end with five epochs. Our source code with every detail will be released, and the satellite images will be available for research purposes only and upon request.

5.1 Cross-view video-based localization

Since there are no existing video-based cross-view localization algorithms, we conduct extensive experiments to dissect the effectiveness and necessity of each component in our framework.

Table 2: Performance comparison on different designs for view projection and sequence fusion (sequence = 4)

		Model	Test-1				Test-2			
		Size	r@1	r@5	r@10	r@100	r@1	r@5	r@10	r@100
View Projection	Ours w/o GVP (Unet)	66.4M	0.08	0.61	1.70	26.24	0.00	0.00	0.00	1.09
	Ours w/o GVP	66.2M	1.66	4.33	7.97	36.35	0.04	0.16	0.20	5.22
Direct Fusion	Conv2D	66.0M	1.25	5.90	10.80	65.91	8.90	18.44	26.61	76.51
	Conv3D	66.0M	15.08	41.57	53.17	93.09	7.00	20.38	30.33	75.90
	LSTM	66.2M	12.53	32.11	50.42	96.93	5.78	15.89	23.01	70.60
Attention based Fusion	Conv2D	66.0M	18.80	47.03	61.75	96.64	11.69	25.03	36.55	81.52
	Conv3D	66.0M	19.65	43.27	58.39	97.41	11.36	24.02	34.45	83.58
	LSTM	66.1M	15.93	47.88	**66.03**	97.61	9.70	24.30	35.26	**85.08**
	Ours	66.2M	**21.80**	**47.92**	64.94	**99.07**	**12.90**	**27.34**	**38.62**	85.00

5.1.1 Geometry-driven view projection.

Although our GVP module is the basis for the following sequence fusion and similarity matching steps, we investigate whether it can be replaced or removed. We first replace it with an Unet and expect the domain correspondences can be learned implicitly during training, denoted as "Ours w/o GVP (Unet)". Next, we remove it from our pipeline and directly feed the original ground-view features to our sequence fusion module, denoted as "Ours w/o GVP". As indicated by the results in Tab. 2, the performance of the two baselines is significantly inferior to our whole pipeline, demonstrating the necessity of our geometry-driven view projection module.

Learned viewpoint-invariant semantic features. To fully understand the capability of our view projection module, we visualize the learned viewpoint-invariant semantic features of our network by using the techniques of Grad-Cam [66]. As seen in Fig. 5, salient features on roads and roads edges are successfully recognized in ground-view images (Fig. 5(a)). The detected salient features in satellite images also concentrate on roads and scene objects along roads edges (Fig. 5(b)). By using our view projection module and the photo-consistency constrained sequence fusion mechanism, the learned global representations of the ground video (Fig. 5(c)) capture similar scene patterns to those of their matching satellite counterparts (Fig. 5(d)).

5.1.2 Photo-consistency constrained sequence fusion.

Our goal is to synthesize an overhead-view feature map from a query ground video. To this end, our PCSF module measures the photo consistency for each overhead view pixel across different ground-view images and fuses them with an attention-based (transformer) architecture. Apart from this design, LSTM (RNNs) and 3D CNNs are also known for their power to handle sequential signals. Hence, we compare with these architectures. For completeness, we also experiment with 2D CNNs.

Direct fusion. We first replace our PCSF module with Conv2D, Conv3D, and LSTM based networks, respectively. The Conv2D-based fusion network takes the projected sequential ground-view features $\mathbf{F}_{t_i}^{p}$ separately and computes the average of the outputs of different time steps. The Conv3D-based fusion network uses its third dimension to operate on the temporal dimension. The LSTM-based network includes two bidirectional LSTM layers to enhance the sequential

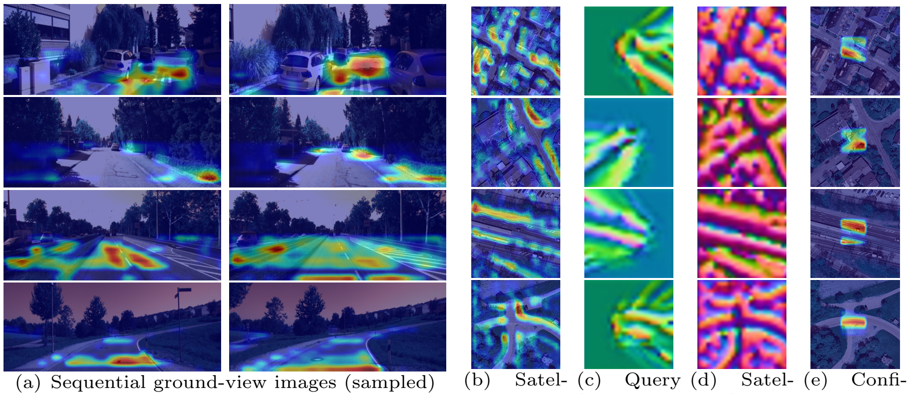

(a) Sequential ground-view images (sampled) (b) Satellite image (c) Query feature (d) Satellite feature (e) Confidence map

Fig. 5: Visualizations of intermediate results of our method. For ground images, the learned activations focus on salient features of the ground and road edges. Interestingly, it automatically ignores the dynamic objects (first row in (a)). The learned activations of satellite images concentrate on the scene objects along the main road (likely visible by a moving vehicle). The fused query video features (c) capture scene objects similar to those of their satellite counterparts (d), and the learned confidence maps attention on the region of road.

relationship encoding. The outputs of the Conv3D-based and the LSTM-based networks are both directly fused features for the query video. The results are presented in the middle part of Tab. 2. It can be seen that the performance is significantly inferior to ours.

Attention-based fusion. Based on the above observations, we infer that it may be difficult for a network to fuse a sequence of features implicitly. Hence, we employ the Conv2D, Conv3D, and LSTM based network to regress the attention weights for the projected features at different time steps, denoted as $\mathbf{N}_{t_i} \in \mathbb{R}^{S \times S}$. Then, the global query descriptor is obtained by a dot product between the attention weight $\mathbf{N}_{t_i}$ and the features $\mathbf{F}^p_{t_i}$. The results are presented in the bottom part of Tab. 2. It can be seen that the attention-based fusion methods all outperform the direct fusion methods, indicating that the attention-based decomposition helps to achieve better performance. Among the attention-based fusion ablations, our method achieves the best overall performance. This should be attributed to the explicit photo consistency computation across different ground-views by our PCSF module.

5.1.3 Different choices for network backbone. In this section, we conduct ablation study on different network backbones, including Vision transformer (ViT) [67], Swin transformer [68], Renet50 [69] and VGG16 [64](ours). Transformers are known of their superior feature extraction ability than CNNs. However, they do not preserve the translational equivariance ability, which however is an essential element in estimating the relative displacement between query camera locations and their matching satellite image centers. Thus, transformers achieve slightly worse performance than CNNs, as indicated by Tab. 3. Compared to VGG16, Resnet50 does not make significant improvement. Hence, following most previous works [2, 6–8, 11], we use VGG16 as our network backbone.

Table 3: Performance comparison with different backbones (sequence = 4)

	Test-1				Test-2			
	r@1	r@5	r@10	r@100	r@1	r@5	r@10	r@100
ViT [67]	20.05	45.13	60.17	97.53	12.86	27.94	**38.86**	81.64
Swin [68]	18.40	47.80	63.73	**99.11**	12.29	22.31	35.29	80.70
Resnet [69]	**22.68**	**55.16**	**67.69**	97.90	9.75	**28.31**	38.45	73.72
VGG16 [64] (Ours)	21.80	47.92	64.94	99.07	**12.90**	27.34	38.62	**85.00**

Table 4: Effectiveness of the scene-prior driven similarity matching (sequence = 4)

	Test-1				Test-2			
	r@1	r@5	r@10	r@100	r@1	r@5	r@10	r@100
Ours w/o SSM	6.35	25.76	41.97	97.61	3.48	9.42	14.03	63.04
Ours w/o U	13.26	36.76	55.72	97.05	10.47	**27.42**	**39.51**	**88.92**
Ours	**21.80**	**47.92**	**64.94**	**99.07**	**12.90**	27.34	38.62	85.00

5.1.4 Scene-prior driven similarity matching.

Next, we study whether the NCC-based similarity matching can be removed. In this experiment, the distance between the satellite features and the fused ground-view features is directly computed without estimating their potential alignments. Instead, they are assumed to be aligned at the satellite image center. The results are presented in the first row of Tab. 4. The performance drops significantly compared to our whole baseline, demonstrating that the network does not have the ability to tolerate the spatial shifts between query camera locations, and our explicit alignment strategy (NCC-based similarity matching) is effective.

Furthermore, we investigate the effectiveness of the learned scene prior by the uncertainty map (Eq. 5). To do so, we remove the term of uncertainty map $\mathbf{U}(\mathbf{F}^s)_{m,n}$ in Eq. (5), denoted as "Ours w/o U". The results in the second row of Tab. 4 indicates the learned uncertainty map boosts the localization performance. Fig. 5(e) visualizes the generated confidence maps (inverse of uncertainty) by our method. It can be seen that the higher confidence regions mainly concentrate on roads, indicating that the confidence maps successfully encode the semantic information of satellite images and recognize the correct possible regions for a vehicle location.

5.1.5 Varying sequence lengths.

One desired property for a video-based localization method is to be robust to various input video lengths after a model is trained. Hence, we investigate the performance of our method on different query video sequences (1-16) using a model trained on sequence 4. Fig. 6 shows that the performance increases elegantly with the increase in number of video sequences. This confirms our general intuition that more input images will increase the discriminativeness of the query place and help boost the localization performance. Note that when keep increasing the sequence length until the cameras at previous time steps exceed the pre-set coverage of the projected features, the performance will not increase but stay same because we did not fuse exceeded information. Scene contents too far from the query camera location are also less useful for localization.

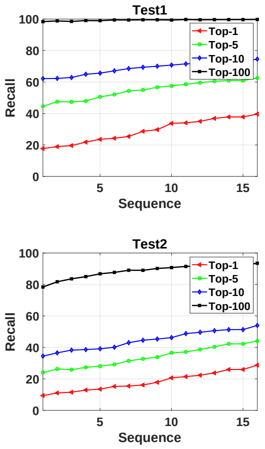
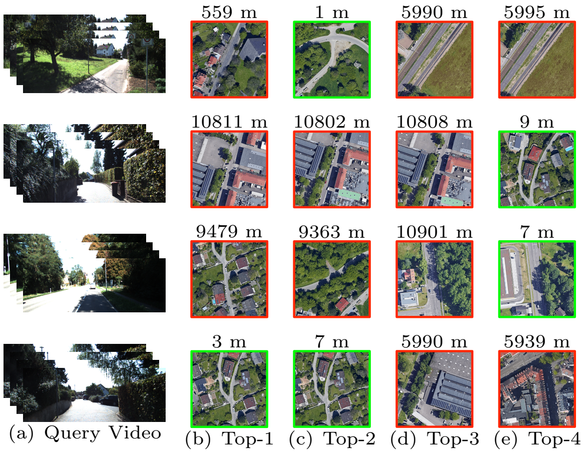


Fig. 6: Recall rates with the increase of input sequence number.



(a) Query Video (b) Top-1 (c) Top-2 (d) Top-3 (e) Top-4

Fig. 7: Qualitative visualization of retrieved results using 4 image frames in a video.

When only using one image for localization, the feature extraction time for a query descriptor is 0.15s. With the increasing of sequence numbers, the query descriptor extraction time increases linearly. We expect this can be accelerated by parallel computation. The retrieval time for each ground image on Test-2 is around 3ms, and the coverage of satellite images in Test-2 is about $710,708$ m^2. It takes 8GB GPU memory when the sequence=4 and 24GB when the sequence=16. We show some qualitative examples of retrieved results in Fig. 7 using sequence number 4.

5.2 Single image-based localization

Single image-based localization is a special case of video-based localization, *i.e.*, when the image frame count in the video is one. In this section, we compare the performance of our method with the recent state-of-the-art (SOTA) that are invented for cross-view single image-based localization, including CVM-NET [2], CVFT [7], SAFA [6], Polar-SAFA [6], DSM [8], Zhu *et al.* [11], and Toker *et al.* [10]. The results are presented in Tab. 5. It can be seen that our method significantly outperforms the recent SOTA algorithms.

Among the compared algorithms, DSM [8] achieves the best performance, because it explicitly addresses the challenge of limited FoV problem of query images while the others assume that query images are full FoV panoramas. By comparing SAFA and Polar-SAFA, we can observe that the polar transform boosts the performance on Test-1 (one-to-one matching) while impairs the performance on Test-2 (one-to-many matching). This is consistent with the conclusion in Shi *et al.* [6] and Zhu *et al.* [11].

Based on SAFA, Zhu *et al.* [11] proposes two training losses: (1) an IoU loss and (2) a GPS loss. However, we do not found the two items work well on the KITTI-CVL dataset. We guess that the IoU loss is only suitable for the panorama case. The limited FoV images in the KITTI-CVL dataset have a

Table 5: Comparison with the recent state-of-the-art on single image based localization

Method	Test-1				Test-2			
	r@1	r@5	r@10	r@100	r@1	r@5	r@10	r@100
CVM-NET [2]	6.43	20.74	32.47	84.07	1.01	4.33	7.52	32.88
CVFT [7]	1.78	7.20	14.40	73.55	0.20	1.29	3.03	16.86
SAFA [6]	4.89	15.77	23.29	87.75	1.62	4.73	7.40	30.13
Polar-SAFA [6]	6.67	17.06	27.62	86.53	1.13	3.76	6.23	28.22
DSM [8]	13.18	41.16	58.67	97.17	5.38	18.12	28.63	75.70
Zhu *et al.* [11]	5.26	17.79	28.22	88.44	0.73	3.28	5.66	27.86
Toker *et al.* [10]	2.79	7.72	11.69	58.92	2.39	5.50	8.90	27.05
Ours	**17.71**	**44.56**	**62.15**	**98.38**	**9.38**	**24.06**	**34.45**	**85.00**

smaller overlap with satellite images than panoramas, and thus the original IoU loss may not provide correct guidance for training. The GPS loss does not help mainly because of the inaccuracy of the GPS data in our dataset. We provide the GPS accuracy analysis of the KITTI dataset in the supplementary material. In contrast, our method does not rely on the accurate GPS tags of ground or satellite images.

5.3 Limitations

Our method assumes that the north direction is provided by a compass, following previous works [3, 6, 8, 10, 11], and the absolute scale of camera translations can be estimated roughly from the vehicle velocity. We have not investigated how significant tilt and roll angle changes will affect the performance, because the tilt and roll angles in the KITTI dataset are very small and we set them to zero. In autonomous driving scenarios, the vehicle-mounted cameras are usually perpendicular to the ground plane. Thus there are only slight changes in tilt and toll during driving.

6 Conclusions

This paper introduced a novel geometry-driven semantic correspondence learning approach for cross-view video-based localization. Our method includes a Geometry-driven View Projection block to bridge the cross-view domain gap, a Photo-consistency Constrained Sequence Fusion module to aggregate the sequential ground-view observations and a Scene-prior driven similarity matching mechanism to determine the location of a ground camera with respect to a satellite image center. Benefiting from the proposed components, we demonstrate that using a video rather than a single image for localization significantly facilitates the localization performance considerably.

Acknowledgments This research is funded in part by ARC-Discovery grants (DP190102261 and DP220100800 to HL, DP220100800 to XY) and ARC-DECRA grant (DE230100477 to XY). YS is a China Scholarship Council (CSC)-funded PhD student to ANU. We thank all anonymous reviewers and ACs for their constructive suggestions.

References

1. Vo, N.N., Hays, J.: Localizing and orienting street views using overhead imagery. In: European Conference on Computer Vision, Springer (2016) 494–509
2. Hu, S., Feng, M., Nguyen, R.M.H., Hee Lee, G.: Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
3. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
4. Regmi, K., Shah, M.: Bridging the domain gap for ground-to-aerial image matching. In: The IEEE International Conference on Computer Vision (ICCV). (2019)
5. Cai, S., Guo, Y., Khan, S., Hu, J., Wen, G.: Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In: The IEEE International Conference on Computer Vision (ICCV). (2019)
6. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. In: Advances in Neural Information Processing Systems. (2019) 10090–10100
7. Shi, Y., Yu, X., Liu, L., Zhang, T., Li, H.: Optimal feature transport for cross-view image geo-localization. In: AAAI. (2020) 11990–11997
8. Shi, Y., Yu, X., Campbell, D., Li, H.: Where am I looking at? joint location and orientation estimation by cross-view matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 4064–4072
9. Zhu, S., Yang, T., Chen, C.: Revisiting street-to-aerial view image geo-localization and orientation estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (2021) 756–765
10. Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L.: Coming down to earth: Satellite-to-street view synthesis for geo-localization. CVPR (2021)
11. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. CVPR (2021)
12. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32** (2013) 1231–1237
13. (https://developers.google.com/maps/documentation/maps-static/overview)
14. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 5297–5307
15. Kim, H.J., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geo-localization. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 3251–3260
16. Liu, L., Li, H., Dai, Y.: Stochastic attraction-repulsion embedding for large scale image localization. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2570–2579
17. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: Proceedings of the IEEE international conference on computer vision. (2017) 3456–3465
18. Ge, Y., Wang, H., Zhu, F., Zhao, R., Li, H.: Self-supervising fine-grained region similarities for large-scale image localization. In: European Conference on Computer Vision, Springer (2020) 369–386
19. Zhou, Y., Wan, G., Hou, S., Yu, L., Wang, G., Rui, X., Song, S.: Da4ad: End-to-end deep attention-based visual localization for autonomous driving. In: European Conference on Computer Vision, Springer (2020) 271–289

20. Castaldo, F., Zamir, A., Angst, R., Palmieri, F., Savarese, S.: Semantic cross-view matching. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2015) 9–17
21. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 891–898
22. Mousavian, A., Kosecka, J.: Semantic image based geolocation given a map. arXiv preprint arXiv:1609.00278 (2016)
23. Tian, Y., Chen, C., Shah, M.: Cross-view image matching for geo-localization in urban environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 3608–3616
24. Hu, S., Lee, G.H.: Image-based geo-localization using satellite imagery. International Journal of Computer Vision **128** (2020) 1205–1219
25. Shi, Y., Yu, X., Liu, L., Campbell, D., Koniusz, P., Li, H.: Accurate 3-dof camera geo-localization via ground-to-satellite image matching. arXiv preprint arXiv:2203.14148 (2022)
26. Zhu, S., Shah, M., Chen, C.: Transgeo: Transformer is all you need for cross-view image geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2022) 1162–1171
27. Elhashash, M., Qin, R.: Cross-view slam solver: Global pose estimation of monocular ground-level video frames for 3d reconstruction using a reference 3d model from satellite images. ISPRS Journal of Photogrammetry and Remote Sensing **188** (2022) 62–74
28. Guo, Y., Choi, M., Li, K., Boussaid, F., Bennamoun, M.: Soft exemplar highlighting for cross-view image-based geo-localization. IEEE Transactions on Image Processing **31** (2022) 2094–2105
29. Zhao, J., Zhai, Q., Huang, R., Cheng, H.: Mutual generative transformer learning for cross-view geo-localization. arXiv preprint arXiv:2203.09135 (2022)
30. Bloesch, M., Omari, S., Hutter, M., Siegwart, R.: Robust visual inertial odometry using a direct ekf-based approach. In: 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE (2015) 298–304
31. Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., Furgale, P.: Keyframe-based visual–inertial odometry using nonlinear optimization. The International Journal of Robotics Research **34** (2015) 314–334
32. Chien, H.J., Chuang, C.C., Chen, C.Y., Klette, R.: When to use what feature? sift, surf, orb, or a-kaze features for monocular visual odometry. 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ) (2016) 1–6
33. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on robotics **32** (2016) 1309–1332
34. Engel, J., Schöps, T., Cremers, D.: Lsd-slam: Large-scale direct monocular slam. In: European conference on computer vision, Springer (2014) 834–849
35. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: 2007 6th IEEE and ACM international symposium on mixed and augmented reality, IEEE (2007) 225–234
36. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: a versatile and accurate monocular slam system. IEEE transactions on robotics **31** (2015) 1147–1163
37. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE transactions on robotics **33** (2017) 1255–1262

38. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics (2021)
39. Mur-Artal, R., Tardós, J.D.: Visual-inertial monocular slam with map reuse. IEEE Robotics and Automation Letters **2** (2017) 796–803
40. Wolcott, R.W., Eustice, R.M.: Visual localization within lidar maps for automated urban driving. 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (2014) 176–183
41. Voodarla, M., Shrivastava, S., Manglani, S., Vora, A., Agarwal, S., Chakravarty, P.: S-bev: Semantic birds-eye view representation for weather and lighting invariant 3-dof localization (2021)
42. Stenborg, E., Toft, C., Hammarstrand, L.: Long-term visual localization using semantically segmented images. In: 2018 IEEE international conference on robotics and automation (ICRA), IEEE (2018) 6484–6490
43. Stenborg, E., Sattler, T., Hammarstrand, L.: Using image sequences for long-term visual localization. In: 2020 International Conference on 3D Vision (3DV), IEEE (2020) 938–948
44. Vaca-Castano, G., Zamir, A.R., Shah, M.: City scale geo-spatial trajectory estimation of a moving camera. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2012) 1186–1193
45. Regmi, K., Shah, M.: Video geo-localization employing geo-temporal feature learning and gps trajectory smoothing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 12126–12135
46. Yousif, K., Bab-Hadiashar, A., Hoseinnezhad, R.: An overview to visual odometry and visual slam: Applications to mobile robotics. Intelligent Industrial Systems **1** (2015) 289–311
47. Scaramuzza, D., Fraundorfer, F.: Visual odometry [tutorial]. IEEE Robotics & Automation Magazine **18** (2011) 80–92
48. Gao, X., Wang, R., Demmel, N., Cremers, D.: Ldso: Direct sparse odometry with loop closure. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE (2018) 2198–2204
49. Kasyanov, A., Engelmann, F., Stückler, J., Leibe, B.: Keyframe-based visual-inertial online slam with relocalization. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE (2017) 6662–6669
50. Liu, D., Cui, Y., Guo, X., Ding, W., Yang, B., Chen, Y.: Visual localization for autonomous driving: Mapping the accurate location in the city maze (2020)
51. Hou, Y., Zheng, L., Gould, S.: Multiview detection with feature perspective transformation. In: European Conference on Computer Vision, Springer (2020) 1–18
52. Hou, Y., Zheng, L.: Multiview detection with shadow transformer (and view-coherent data augmentation). In: Proceedings of the 29th ACM International Conference on Multimedia. (2021) 1673–1682
53. Vora, J., Dutta, S., Jain, K., Karthik, S., Gandhi, V.: Bringing generalization to deep multi-view detection. arXiv preprint arXiv:2109.12227 (2021)
54. Ma, J., Tong, J., Wang, S., Zhao, W., Zheng, L., Nguyen, C.: Voxelized 3d feature aggregation for multiview detection. arXiv preprint arXiv:2112.03471 (2021)
55. Zhang, Q., Lin, W., Chan, A.B.: Cross-view cross-scene multi-view crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 557–567
56. Zhang, Q., Chan, A.B.: Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 8297–8306

57. Zhang, Q., Chan, A.B.: 3d crowd counting via multi-view fusion with 3d gaussian kernels. In: Proceedings of the AAAI conference on artificial intelligence. Volume 34. (2020) 12837–12844
58. Zhang, Q., Chan, A.B.: Wide-area crowd counting: Multi-view fusion networks for counting in large scenes. International Journal of Computer Vision (2022) 1–23
59. Chen, L., Sima, C., Li, Y., Zheng, Z., Xu, J., Geng, X., Li, H., He, C., Shi, J., Qiao, Y., et al.: Persformer: 3d lane detection via perspective transformer and the openlane benchmark. arXiv preprint arXiv:2203.11089 (2022)
60. Shi, Y., Campbell, D.J., Yu, X., Li, H.: Geometry-guided street-view panorama synthesis from satellite imagery. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
61. Shi, Y., Li, H.: Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 17010–17020
62. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4104–4113
63. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. (2017) 5998–6008
64. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
65. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
66. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017)
67. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
68. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 10012–10022
69. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778