This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Unconstrained Cross-View Pose Estimation

Alexander Wollam[1] Kyle Ashley[2] Maxim Shugaev[2] Oliver Arend[2]
Ilya Semenov[2] Hadis Dashtestani[2] Sumved Ravi[2] Nathan Jacobs[1]
[1]Washington University in St. Louis [2]Blue Halo

Abstract

Cross-view pose estimation entails predicting the relative 3 Degrees-of-Freedom (3DoF) pose of an image within an aerial view. Existing work focuses on imagery in controlled settings featuring highly constrained parameters. In contrast, a wide variety of camera parameterizations are seen in-the-wild across tasks where such estimation is useful. To address this gap, we propose a method capable of performing cross-view pose estimation in these less constrained scenarios with ground-view images of unknown FoV, pitch, roll, and projection type (panoramic or rectilinear). Namely, our method avoids common assumptions—such as gravity/horizon alignment needed for geometric-based projections—and purely relies on a transformer to learn the cross-view relationships in a data-driven manner, paired with prediction modules to enable continuous querying of the pose search space. Evaluations of our approach demonstrates it's ability to perform competitively with the state-of-the-art over the VIGOR benchmark, while maintaining performance in those harder less constrained scenarios. This supports our work as the first generalized approach to this task that is capable of operating with less-constrained imagery.

1. Introduction

Cross-view pose estimation lies at the intersection of image pose estimation and cross-view image geo-localization and as such has many applications. While most existing literature for cross-view image geo-localization is primarily motivated through the lens of determining where in the world an image was captured, newer work shifts this approach to predict and leverage orientation as well [25]. This concept has since extended to the cross-view pose estimation task with recent work motivated through the lens of tasks such as autonomous driving, robotics, and augmented reality [27, 37, 45], where precise pose of ground-level cameras is crucial and many camera parameters are predetermined and consistent. In practice, there are more applications that are less explored, such as an extension of cross-view geo-localization at a more fine-grained level, as an augmentation step continuing [25] to improve geo-localization, as well as towards a more generic model capable of operating across diverse data conditions and/or camera parameterizations.

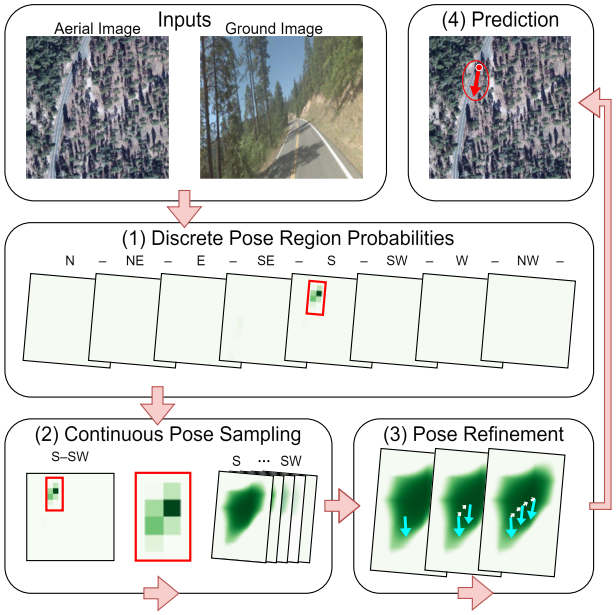

Figure 1. An overview of our pose prediction pipeline (Sec. 3.1). Our model first predicts global pose region probabilities (1) to reduce the search space and avoid local minima. Then we sample poses constrained by this and query the model for their alignment scores (2). Lastly, we optimize the best pose via gradient ascent (3) to get a fine-grained pose prediction (4). Aerial image is from CVUSA [42], sourced from Bing Maps © Microsoft Corporation, ground image from CVUSA [42], sourced from Google Street View © Google, Inc.

In contrast, existing work focuses on settings where imagery is constrained by assumptions of known/consistent camera parameters, which limit their applicability to diverse environments and settings. For instance, fine-grained cross-view geo-localization methods often rely on geometric transformations, such as polar transformations [25] or homography estimation [23, 37, 40], to align ground and satellite images, reducing the problem to one of image

alignment. These methods have shown impressive accuracy in cross-view pose estimation in their constrained problem settings, but haven't been tailored to less structured or more variable settings where such assumptions may break down.

Recent advanced methods provide a way to get around some of these limitations. Specifically, transformer-based architectures, known for their ability to capture long-range and complex relationships, have demonstrated high performance on various vision tasks. In cross-view pose estimation, transformer-based models offer the potential to leverage existing large datasets to learn the relationships between ground and aerial images without the need for explicit geometric projections, making them more adaptable to diverse datasets and scenarios.

In this paper, we propose a transformer-based approach which implicitly learns cross-view relationships without geometric constraints to avoid reliance on specific camera parameterizations. Our method allows for flexible estimation that can handle diverse environments and images. Additionally, our prediction modules enable searching the pose space in a continuous manner, allowing for arbitrarily fine-grained pose extraction.

Our contributions are threefold:

- We propose a novel approach that leverages a transformer-based model, enabling accurate cross-view pose estimation in poorly constrained environments.
- We introduce a pair of prediction modules that enables continuous querying of poses at arbitrarily fine resolution, improving training and accuracy.
- We demonstrate the effectiveness of our method over existing baselines and towards more diverse settings through the use of rotated gnomonic/rectilinear projections, achieving competitive results in controlled environments while taking the first steps towards high performance in more general scenarios.

2. Related Work

Cross-view image retrieval: This task approaches image geo-localization from the perspective of matching query ground images to a database of satellite images. Since its inception, it has seen much work in dataset creation and with different assumptions, enabling many different approaches [15, 16, 32, 41, 42, 52]. In general, the task involves creating a descriptor of the query ground image that is then matched to a descriptor for each candidate aerial patch [10, 17, 24, 47]. While initial attempts at this task [15, 42] achieved reasonable results, approaches have since improved to better deal with large appearance and perspective gaps. Polar transforms have become a widely leveraged tool that works to close this perspective gap to improve performance [24], and this has been combined with the use of orientation considerations and newer deep architectures [47, 53]. Additionally, synthesis from one view to the other has also been explored as a way to close these gaps [14, 21, 26, 33]. A limitation faced by these approaches is the common implicit assumption that the ground image is located near the center of the aerial image which may not be true in general. There has been some work to loosen this assumption, such as with [52] where they propose a dataset featuring images with variable alignment, however there is still much work to be done for these cases.

Cross-view camera pose estimation: Beyond image-level cross-view localization through retrieval, cross-view camera pose estimation represents a finer-scale extension that looks to determine the 3DoF pose of a ground image within an aerial patch. Initial work in extending to variable alignments was performed by [52] introducing the VIGOR benchmark. Most approaches leverage projections into satellite space from which matching-based optimization can take place [5, 13, 23, 26, 27, 37, 38, 40, 44, 45], however the specific method in which this is done varies. One major approach is to leverage Birds Eye View (BEV) projections of the ground view, and then matching that across different candidate poses within the satellite patch [5, 22, 27]. While the above approaches have been optimized over denser urban environments, there has also been some recent work in harder, rural scenes using BEVs as well [12]. A downside of these BEV-based approaches, however, is that performing a dense search of candidate poses becomes expensive. Other approaches accelerate this [13, 40, 45] by aggregating features into a single vector [45], pre-computing masks [13], and reformulating the matching procedure as homography estimation [40]. Alternatively, keypoints have also been used to refine pose matching by detecting and projecting them into the satellite space [37, 38], achieving high levels of localization. While these approaches have demonstrated much success, there is still work to be done in extending to more inconsistent/unknown camera intrinsics/extrinsics.

Transformers in Multi-View Vision: Transformers, first introduced by Vaswani [35] have become an instrumental tool used across computer vision tasks [8]. Early foundational work applying transformers to the vision domain [4] demonstrated the potential of this architecture towards vision, with many additional formulations following [7, 18].

Leveraging these general approaches, transformers have been utilized successfully in a variety of multi-view settings. These include cross-view tasks [47, 50, 53], multi-view reconstruction/stereo [2, 3, 36, 39, 48, 51], video-related tasks [1, 19, 20, 46], and multi-view object pose estimation [9, 28, 49] among others. In particular, multiview camera pose estimation through reconstruction objectives have also demonstrated the ability of transformers to implicitly merge multi-view features without prior pose knowledge [48]. These above approaches have thus demonstrated the ability of transformers to capture important multi-view relationships, and motivate applying them to our task.

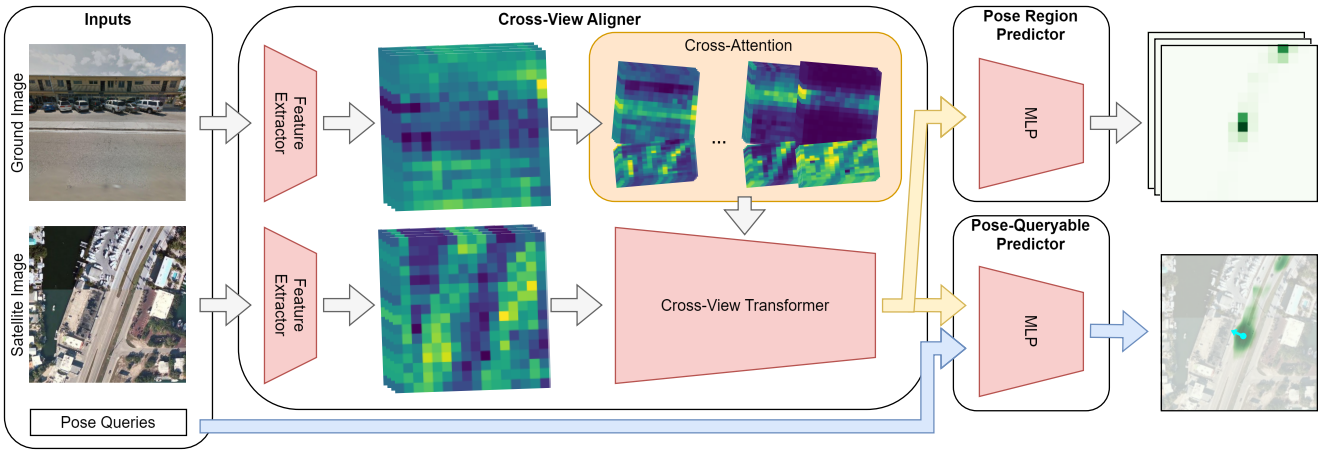

Figure 2. An overview of our proposed cross-view pose estimation method. It consists of three sections: a cross-view aligner, pose region predictor and pose-queryable predictor. A sample input and output are provided to visualize predictions. Aerial image is from CVUSA [42], sourced from Bing Maps © Microsoft Corporation, ground image from CVUSA [42], sourced from Google Street View © Google, Inc.

3. Method

This paper proposes a novel approach to cross-view 3DoF pose estimation that is robust to diverse and unknown camera parameters. This design, illustrated in figure 2, is composed of a base cross-view transformer (sec 3.2), a pose-queryable predictor (sec 3.4), and a discrete pose region predictor (sec 3.3). An overview of how these components work together during training time (sec 3.5), and in inference time (sec 3.1) is also provided.

3.1. Inference: Pose Prediction Pipeline

During inference time, we predict the pose using a three step process (figure 1): (1) Discrete Pose Region Selection, (2) Continuous Pose Sampling, and (3) Pose Refinement.

Discrete Pose Region Selection: In this first stage, the Pose Region Predictor (Sec. 3.3) produces output probabilities for each predefined region of the pose space which is used to reduce the search space in a single step, avoiding a comprehensive search of the entire space which may otherwise be expensive or struggle with local minima. Then, we simply select the top predicted region.

Continuous Pose Sampling: In this second stage, we sample a number of poses within the selected region in parallel, predict their scores using the Pose-Queryable Predictor (Sec. 3.4), then select the pose with the best score. See Sec. A.4 in the supplementary for additional details.

Pose Refinement: In this last stage, we take the pose with the highest score from the second stage, then optimize over the pose by maximizing the score through gradient ascent. Finally, we record the scores throughout this optimization process and select the pose with the highest score. See Sec. A.4 in the supplementary for additional details.

3.2. Cross-view Transformer

For our task, the model takes in a satellite and ground-level image that lacks meta-data and encodes each using a standard image feature extractor. For computational reasons, we chose a lightweight Swin transformer [18]. Then, we use the satellite feature patches as input to the cross-view transformer; cross-attention is then performed in each layer with the ground image features. Finally, the output is trained using our predictors to encode an implicit similarity map across the satellite view pose space.

3.3. Pose Region Predictor

For prediction, we use two simple MLP-based modules. For this first Pose Region Predictor module, each output patch of our transformer is run through an MLP to predict a set of region scores for that patch's spatial region, one score for each predetermined orientation bin. By using softmax over all scores, we get a discrete pose-region probability distribution over the entire space. This can be leveraged in inference and training time (Sec. 3.1 and Sec. 3.5).

3.4. Pose-Queryable Predictor

For this Pose-Queryable Predictor module, we enable querying pose scores differentiably over the pose space using two simple components: a pose-conditioned MLP and bilinear interpolation-based aggregation.

Pose-Conditioned MLP: In this component, it takes as input the output features of a given patch and the queried pose's orientation and relative location w.r.t. the patch's center location, then it predicts a candidate score using an MLP. Formally, given a patch feature f_p, relative location $loc_{x,y}$, and orientation ori, we predict a candidate score via

$$score = MLP(f_p, loc_{x,y}, cos(ori), sin(ori)).$$

Bilinear Interpolation-Based Aggregation: As the candidate scores above may differ given different feature inputs, and the output patch features themselves will more accurately predict for poses spatially closer to them, to predict the overall score, we select and aggregate candidate scores according to bilinear interpolation. Namely, by interpreting each candidate score as a value located at their respective base patch's center locations, bilinear interpolation of these values at the pose's location entails performing a weighted mean of the closest patch's scores. See Sec. A.3 of the supplementary for more details.

In contrast to the naive approach of simply selecting the feature associated with the patch that overlaps the pose and using its score—which suffers from discontinuity/non-differentiability in output scores across patch boundaries—this strategy enables full differentiability in output scores across the full space, allowing for the pose refinement from Sec. 3.1.

3.5. Training

During training of our model, we use a set of losses to train our two predictors. For the Pose-Queryable Predictor, we use a contrastive loss to maximize the score at the true pose. For the Pose Region Predictor, we use a standard Cross-Entropy loss to maximize the score for the region containing the true pose. Finally, we define a consistency loss to ensure the two predictors are largely consistent. Altogether, the overall loss function is given as:

$$loss = \lambda_1 L_{contrast} + \lambda_2 L_{CE} + \lambda_3 L_{consistency} \tag{1}$$

3.5.1. Losses

Contrastive loss: Here, we use the infoNCE loss [34]. Given a set of pose scores, s, a ground truth pose score s_{gt} and a temperature parameter τ, this loss is defined as:

$$L_{contrast} = -\log\left(\frac{\exp\left(\frac{s_{\text{gt}}}{\tau}\right)}{\sum_i \exp\left(\frac{s_i}{\tau}\right)}\right) \tag{2}$$

Here, the ground-truth pose score is maximized with respect to negative pose scores. To sample negative poses to contrast with we sample according to two distributions. The first is a uniform distribution across the entire pose space, ensuring the model minimizes everywhere. The second is according to the Pose Region Predictor's output regional probability distribution. Here, we sample regions according to this distribution, then uniformly within those regions. This enables sampling from harder regions where poses have higher scores to get more useful negatives in our training. Since querying poses is inexpensive in our model, we sample thousands of negative poses per example, without significant overhead, to accelerate training.

Region Cross Entropy loss: For the Pose Region Predictor, we leverage the standard cross-entropy loss over the softmax of the output logits. Given output region logits z and the true region label y, this is defined as:

$$L_{CE} = -\sum_i y_i \log\left(\frac{\exp(z_i)}{\sum_j \exp(z_j)}\right) \tag{3}$$

Consistency loss: While the Pose-Queryable Predictor and Pose Region Predictor will largely agree, this alone does not guarantee close distributional alignment which is important for use in training and inference, incentivizing this loss. For the consistency loss, we need to transform the two outputs into similar spaces. Specifically, we first take the queried pose scores and aggregate them into their corresponding regions. Then, we minimize its difference with the direct region predictions by maximizing their cosine similarity. Explicitly, let Y_{agg} be the region-aggregated queried-pose-score vector and Y_{reg} be the direct region predicted vector, where each vector component is a region score. Then the consistency loss function is given by:

$$L_{consistency} = -\frac{Y_{\text{agg}} \cdot Y_{\text{reg}}}{\|Y_{\text{agg}}\| \, \|Y_{\text{reg}}\|} \tag{4}$$

We leverage cosine similarity since the vectors' will have slightly different distributions. This is because they are optimized slightly differently and the aggregation can lower the correct region's aggregated score due to the randomness of sampling and since other poses in the correct region will tend to be negative; hence, values won't perfectly match.

4. Experiments

In this section, we introduce the datasets used for evaluation, the evaluation metrics, and our implementation details. Then, we discuss our performance with respect to existing approaches, before expanding to more general settings.

4.1. Datasets

The two datasets we select for evaluations are CVUSA [42] and VIGOR [52]. These two were selected in part due to their use of panoramas for their ground-view images, in contrast to KITTI [6] and Ford-AV [29] which are other often-used datasets. With Panoramas, we can explore varied projection types and camera parameterizations through the use of rectilinear projections to produce standard camera imagery. By interpreting the panorama as a sphere of rays around the camera, this projection allows us to project a subset onto a tangent plane, mimicking the way images are collected by most cameras. By doing this, and varying the resulting projection's pitch, yaw, and roll, we can explore less constrained imagery.

CVUSA Dataset [42] contains geo-tagged matching aerial and ground-level panoramas collected uniformly across the US. In this dataset, all satellite images are approximately center-aligned with the panoramas. To make

Orientation	Method	Same-Area				Cross-Area			
		↓Localization (m)		↓Orientation (°)		↓Localization (m)		↓Orientation (°)	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
Known	CVR [52]	8.82	7.68	-	-	9.45	8.33	-	-
	CVML [44]	9.86	4.58	-	-	13.06	6.31	-	-
	SliceMatch [13]	5.18	2.58	-	-	5.53	2.55	-	-
	CCVPE [45]	3.60	1.36	-	-	4.97	1.68	-	-
	Boosting [27]	4.12	1.34	-	-	5.16	1.40	-	-
	HC-Net* [40]	2.65*	1.17*	1.92*	1.04*	3.35*	1.59*	2.58*	1.35*
	LDFF [30]	3.03	**0.97**	-	-	5.01	2.42	-	-
	C2F-CCPE [31]	3.17	1.34	-	-	3.94	1.68	-	-
	FG^2 [43]	**1.95**	1.08	-	-	**2.41**	**1.37**	-	-
	Ours	3.66	1.46	-	-	4.65	1.63	-	-
Unknown	SliceMatch [13]	6.49	3.13	25.46	4.71	7.22	3.31	25.97	4.51
	CCVPE [45]	3.74	1.42	12.83	6.62	5.41	1.89	27.78	13.58
	HC-Net* [40]	5.87*	2.18*	48.68*	2.28*	6.86*	2.59*	52.69*	2.89*
	LDFF [30]	4.97	1.90	11.20	1.59	7.67	3.67	17.63	2.94
	C2F-CCPE [31]	**3.55**	**1.41**	11.58	5.02	**4.75**	**1.83**	16.57	4.95
	FG^2 [43]	3.78	1.70	12.63	**1.44**	5.95	2.40	28.41	2.20
	Ours	4.31	1.56	**6.57**	1.52	5.84	1.87	**9.26**	**2.11**

Table 1. Comparison of localization and orientation across different regimes on the VIGOR [52] dataset evaluated using only the easier *positive aerial views*. **Best in Bold**. We use the same model checkpoint across orientation regimes. Note "*" indicates results that use the different location labels by Wang et al. [40] and also always optimizes orientation; for the unknown regime, we evaluate using Wang et al. [40]'s provided checkpoints and evaluation code and implement their proposed BEV-rotation scheme.

Orientation	Method	Same-Area				Cross-Area			
		↓Localization (m)		↓Orientation (°)		↓Localization (m)		↓Orientation (°)	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
Known	CVML [44]	13.45	5.39	-	-	17.13	7.78	-	-
	Ours	**4.61**	**1.60**	-	-	**5.92**	**2.00**	-	-
Unknown	CCVPE* [45]	8.32	**1.61**	12.87	5.03	14.71	2.63	26.97	8.22
	Ours	**5.40**	1.76	**7.87**	**1.68**	**7.42**	**2.19**	**11.11**	**2.24**

Table 2. Comparison of localization and orientation on VIGOR [52] evaluated with the more difficult *semi-positive aerial views* included. **Best in Bold**. This regime includes matches near edges of the aerial view, in contrast to the positive-only regime which only matches in the center quarter locations. We use the same checkpoints as in table 1 and retrain and evaluate CCVPE* for an unknown baseline.

compatible with cross-view pose estimation, we use their satellite images collected at zoom-level 18, from which we crop down to 19 so we can leverage random crops to simulate location misalignment. We choose this dataset for our evaluations as it is one of the more comprehensive cross-view datasets, featuring ∼1 million cross-view pairs over a relatively diverse range of regions in the US. In contrast to existing work which leverages much smaller datasets that are more geographically and semantically constrained, CVUSA provides the data quantity and diversity that better supports our goals towards generalization. Importantly, as our architecture is transformer-based, smaller datasets will easily overfit [11] thus requiring a larger dataset for training. Details of the evaluations are covered in section 4.5.

VIGOR Dataset [52] contains ground-level panoramas with associated geo-tags and corresponding aerial images collected in four cities in the US. Each aerial image corresponds to a spatial resolution of around 70m × 70m, and four of them match to each ground-level panorama. Of these four, 1 is positive and the other 3 are semi-positive, which are defined such that the ground-level image is located in the center quarter of the positive aerial views, but not the semi-positive ones. Additionally, Lentsch et al. [13] and following them, Wang et al. [38] noted errors in the labels which they both provide a different correction for. As there are more published results available for Lentsch et al. [13]'s corrected labels, we follow this approach for better consistency to other work. In contrast to existing approaches which train and evaluate over only the positive aerial views, we use both positive and semi-positives as our approach is not directly hindered by location alignments near aerial image edges which may otherwise cause difficulty with other approaches, and doing so retains consistency with our work with CVUSA. For the semi-positives,

				VIGOR Same-Area (zoom 20)				CVUSA (zoom 19)			
				↓Localization (m)		↓Orientation (°)		↓Localization (m)		↓Orientation (°)	
Projection	H-FoV	Pitch	Roll	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Panoramic	Full*	±0°	±0°	5.40	1.76	7.87	1.68	7.57	3.13	4.51	1.55
	180°	±0°	±0°	5.86	1.82	8.72	1.73	14.02	4.19	7.97	2.87
		±10°		6.13	1.85	8.86	1.77	14.30	4.53	9.40	3.08
	90°	±0°	±0°	6.04	1.88	8.89	1.76	14.73	4.59	9.50	2.97
		±10°		6.39	1.92	9.10	1.77	15.42	4.82	10.71	3.03
Gnomonic	90°	±0°	±0°	9.12	2.81	13.40	2.63	21.02	6.54	14.62	3.92
		±10°	±0°	9.62	3.53	16.10	2.76	22.64	7.75	18.02	4.12
		±0°	±20°	9.18	3.22	16.16	2.63	22.73	7.93	17.19	4.16
		±10°	±20°	10.16	4.28	17.42	3.00	23.83	8.29	19.47	4.17
	60°	±0°	±0°	9.87	3.70	17.07	2.62	21.55	7.71	14.77	3.93
		±10°	±0°	9.89	3.98	17.09	2.80	23.34	7.89	18.02	4.15
		±0°	±20°	10.11	4.38	17.82	2.77	23.89	8.20	18.33	4.17
		±10°	±20°	10.45	4.62	18.50	3.15	24.17	8.34	19.56	4.19
	30°	±0°	±0°	10.70	5.20	18.09	2.79	24.27	8.31	17.34	3.93
		±10°	±0°	10.76	5.38	18.66	3.18	25.22	8.71	19.33	4.17
		±0°	±20°	10.86	5.24	19.22	3.16	25.29	9.22	19.49	4.19
		±10°	±20°	11.08	5.67	19.71	3.54	25.26	9.34	20.38	4.21

Table 3. Performance over VIGOR [52] same-area split with semi-positives and CVUSA [42] across image projections, horizontal FoVs, camera pitch, and camera roll. VIGOR performance evaluations are all done on the *same checkpoint*, as are CVUSA's. Note "*" indicates different behaviour between CVUSA and VIGOR, since CVUSA'a maximum horizontal-FoV is ∼332° and VIGOR's is 360°; this section evaluates at that maximum.

using the corrected location labels reveals some of these are not aligned with their respective ground images; we filter these out. By extending to semi-positives, this allows us to demonstrate our performance in a harder matching regime where important features may be 'out-of-view,' which is important for the more general settings where prior alignments are poor or non-existent, though we evaluate both regimes to compare with previous work. Finally, we follow convention and use the same-area and cross-area splits from [52] for comparison with existing approaches.

4.2. Implementation Details

The spatial size of the ground and satellite images are 256 x 256 in both datasets. We encode them using lightweight custom Swin transformers [18] for computational reasons. After being encoded, this results in feature maps of 16 x 16 and a channel size of 256. For the cross-view transformer in section 3.2, we use 8 layers. For both predictors in sections 3.4 and 3.3, we use simple 2-layer MLPs and use 8 orientation bins in the pose-region predictor, dividing orientation space into 45° sections. During training, our model is trained with a learning rate of 1×10^{-4}, and we use a batch size of 120; for our contrastive loss we sample 2,000 negative poses per example. For VIGOR, we pre-train on CVUSA before finetuning on VIGOR to prevent overfitting.

4.3. Evaluation Metrics

For our evaluation metrics, we follow the existing convention used for the VIGOR benchmark of reporting the mean and median error in meters between the predicted and ground truth location over all test image pairs. Similarly, for orientation prediction, we report the mean and median absolute angular difference between the predicted and ground truth orientation in degrees. For consistency, we continue these choices for our in-depth evaluation over both VIGOR and CVUSA across image parameterizations.

4.4. Baseline Comparison

For our baseline comparisons, seen in table 1 and table 2, we compare our approach to previous work on the Same-Area and Cross-Area splits of the VIGOR benchmark. We also evaluate in both the positive-only setting and in the positive + semi-positive setting and provide baselines as they exist.

Same-Area Pose Estimation: In the known orientation regime, the problem reduces to the easier secondary task of 2DoF location estimation. Under this regime, with positives only, we see that our model performs similar to CCVPE [45] and Boosting [27], but is worse than the best LDFF [30] and FG^2 [43]. For completeness we also compare to CVML [44] with semi-positives present as well, and demonstrate superior results, though it is slightly less performant than over positives-only.

In the more difficult primary task with unknown orientation, we see that with positives-only, on localization, our approach performs similarly to CCVPE [45] and LDFF [30] but is worse than the best C2F-CCPE [31]. On orientation however, our method performs the best by a significant margin for most metrics, though FG^2 [43] has close me-

dian performance. When including semi-positives, we train CCVPE [45] and compare to it, as it performs the best overall out of the baselines with released code, but our method is superior for most metrics.

Cross-Area Pose Estimation: In the 2DoF known orientation regime, we see similar trends. With positives-only, we perform similarly to CCVPE [45] and Boosting [27] in localization, achieving marginally better than CCVPE [45] and better than Boosting [27] with respect to mean, but worse in median. We, however, perform worse than the best FG^2 [43]. For completeness, in the regime with semi-positives, we demonstrate superior results to CVML [44].

In the harder 3DoF unknown orientation setting, using only positives, we perform similar to CCVPE [45] and FG^2 [43] in localization, but are worse than the best C2F-CCPE [31]. For orientation, we achieve the best mean and median results. Again, we train and compare to CCVPE [45] in the semi-positive setting, but achieve the best results for all metrics.

Qualitative Discussion: Overall, we note a few key points. For one, our approach maintains a competitive level of performance to the existing work with respect to localization, while exhibiting notably superior orientation ability when no prior is given. Specifically, our mean orientation error tends to be about half the next best result. This likely is because our predictor allows for search and optimization with respect to orientation in the same way it does for localization, whereas other work such as CCVPE [45] and SliceMatch [13] discretize the orientation space for matching. While its localization is generally slightly worse than the top models with respect to localization, this is due to our model being designed to not leverage any prior geometric or pixel-based constraints which otherwise can aid fine-grained localization, with the benefit that it performs significantly better in orientation. Additionally, our performance with semi-positives is largely maintained, despite it being more difficult.

		↓Localization (m)		↓Orientation (°)	
Dataset	Method	Mean	Median	Mean	Median
VIGOR	CCVPE*	8.01	**1.67**	13.22	5.33
	Ours	**5.40**	1.76	**7.87**	**1.68**
CVUSA	CCVPE*	16.56	3.68	51.43	45.68
	Ours	**7.57**	**3.13**	**4.51**	**1.55**

Table 4. Performance over VIGOR [52] same-area split with semi-positives and CVUSA panoramas. Evaluations are done on the *same checkpoint* for each method. *CCVPE is retrained in this setting.

4.5. Effects of Image Parameterization

For our evaluations using CVUSA [42] and VIGOR [52] with variably parameterized images, we explore projection type, FoV, pitch, and roll (table 3). Here, we compare our results to CCVPE retrained in these settings as it was the best overall performing model with available code. We use the same checkpoint across VIGOR evaluations; same for CVUSA. For VIGOR, we utilize the same-area split with semi-positives. For CVUSA (see Sec. 4.1) we use satellite images with random alignment at zoom-level 19 in contrast to VIGOR at 20. As such, CVUSA performance is worse in expectation, as the pose search space is then four times as large in the CVUSA setting and has lower spatial resolution. Indeed, table 4 shows performance is stronger on VIGOR for both methods, with ours performing better overall.

		↓Localization (m)		↓Orientation (°)	
H-FoV	Method	Mean	Median	Mean	Median
360	CCVPE*	8.01	**1.67**	13.22	5.33
	Ours	**5.40**	1.76	**7.87**	**1.68**
180	CCVPE*	13.53	3.73	24.03	7.57
	Ours	**5.86**	**1.82**	**8.72**	**1.73**
90	CCVPE*	21.12	14.11	43.16	16.14
	Ours	**6.04**	**1.88**	**8.89**	**1.76**

Table 5. Performance over localization and orientation over VIGOR [52] same-area split with semi-positives over different horizontal FoV panoramic cutouts. Evaluations are all done on the *same checkpoint* for each method. *CCVPE is retrained in this setting.

		↓Localization (m)		↓Orientation (°)	
H-FoV	Method	Mean	Median	Mean	Median
90	CCVPE*	15.36	4.74	37.72	8.95
	Ours	**9.12**	**2.81**	**13.40**	**2.63**
60	CCVPE*	16.86	7.62	50.83	14.48
	Ours	**9.87**	**3.70**	**17.07**	**2.62**
30	CCVPE*	18.14	9.82	60.72	24.52
	Ours	**10.70**	**5.20**	**18.09**	**2.79**

Table 6. Performance over localization and orientation over VIGOR [52] same-area split with semi-positives over rectilinear projections of different horizontal FoVs. Evaluations are all done on the *same checkpoint* for each method. *CCVPE is retrained in this setting.

Projection Type: We explore two image projection types: panoramic and gnomonic, as described in section 4.1. Our results over these projection types show that the model makes reasonable pose predictions in all cases, however the gnomonic images tend to localize and orient worse even at the same parameterization. Specifically, we see a $\sim$1 meter median localization increase for VIGOR and a $\sim$2 meter increase for CVUSA; similarly we see a $\sim 1^\circ$ median orientation increase for both. This performance drop may indicate that panoramic-crops may have features that enable slightly better predictions than gnomonic, such as feature curvature.

Similarly between tables 5, 6 the panoramic images tend to have higher performance, though CCVPEs struggle at the lower end of its FoV training/eval range.

Field of View: For the FoV, we specifically record the effects on performance with respect to horizontal FoV. This was done, since performance was not seen to vary significantly across reasonable vertical FoVs, intuitively since alignment performance is impacted by how much of the scene is visible which is dictated horizontally for images that aren't sideways. In the results of table 3, we see the expected trend that as horizontal FoV decreases, all metrics correspondingly decrease. For both VIGOR and CVUSA the median localization error increases by $\sim$3 when comparing the full-FoV panoramas to the 30° gnomonic images. For orientation error, the median over VIGOR increase by $\sim$1.6 and CVUSA $\sim$2.5. This trend holds within projection types, pitches, and rolls as well as independently. When comparing to CCVPE in tables 5, 6, we see that our model holds its performance significantly better at lower FoVs.

		↓Localization (m)		↓Orientation (°)	
Pitch	Method	Mean	Median	Mean	Median
±0°	CCVPE*	15.36	4.74	37.72	8.95
	Ours	**9.12**	**2.81**	**13.40**	**2.63**
±10°	CCVPE*	16.98	7.93	49.27	14.58
	Ours	**9.62**	**3.53**	**16.10**	**2.76**

Table 7. Performance over VIGORsame-area split with semi-positives with 90° horizontal FoV projections across different pitch noise levels. Evaluations are all done on the *same checkpoint* for each method. *CCVPE is retrained in this setting.

Pitch: For pitch, we vary this parameter by up to ±10°. No evaluation is done with the full-fov panoramas, as pitch does not change what information is visible. In evaluation, we see that varying pitch from being horizon-aligned results in slightly worse performance; this is seen by a similar amount across projections, FoVs and rolls. For CVUSA, median localization decreases by $\sim$0.2 meters and median orientation by $\sim$0.1° in aggregate. For VIGOR, median localization decreases by $\sim$0.03 meters for panoramas and typically $\sim$0.2 meters for gnomonic, with a couple exceptions that are $\sim$1 meter worse; for median orientation panoramas decrease by $\sim$0.03° and gnomonic by $\sim$0.3° in aggregate. Overall, with ±10° of pitch, there is some drop in performance but it is relatively minor. When comparing to CCVPE in table 7, we see that our performance holds better as pitch varies.

Roll: For roll, we vary this up to ±20° and evaluate over only gnomonic images; we again see a behaviour of worsening performance when rolled. For both VIGOR and CVUSA, we get a $\sim$0.6 meter median localization drop with roll, and a $\sim$0.2° median orientation decrease. Overall, we again see a minor performance drop. Again, table 8 demonstrates that our method maintains performance better.

		↓Localization (m)		↓Orientation (°)	
Roll	Method	Mean	Median	Mean	Median
±0°	CCVPE*	15.36	4.74	37.72	8.95
	Ours	**9.12**	**2.81**	**13.40**	**2.63**
±20°	CCVPE*	17.27	8.19	51.55	15.61
	Ours	**9.18**	**3.22**	**16.16**	**2.63**

Table 8. Performance over localization and orientation over VIGORsame-area split with semi-positives with 90° horizontal FoV projections across different roll noise levels. Evaluations are all done on the *same checkpoint* for each method. *CCVPE is retrained in this setting.

Qualitative Discussion: Across the evaluations done, we can see that our approach is successful at handling variable projections, FoVs, pitch, and roll, though is not invariant to them. Notably, decreasing horizontal FoV has the greatest impact by decreasing the number of features available to leverage for prediction, making matching more ambiguous. Varying pitch and roll similarly effected performance, but to a lesser extent. Despite this, the median performance of the model is still relatively high even under harder parameterizations and is less effected than the baseline. Additionally, when moving to the harder CVUSA scenes, performance drops but the model is still able to make reasonable predictions in most cases.

5. Conclusion

In this paper we introduced a novel approach to cross-view pose estimation that uniquely handles unknown image projections and parameterizations. By leveraging learnable attention mechanisms, our model generalizes to less constrained settings not otherwise possible with rigid geometric constraints. Through our implicit latent matching, our approach lends itself to both pose region prediction, as well as to a continuous queryable prediction method, allowing for fast pose optimization on an arbitrarily fine scale. Our approach achieves 3DoF cross-view pose estimation performance that is competitive with the state-of-the-art on VIGOR and maintains much of this performance in harder settings. It can also leverage prior pose knowledge by directly incorporating such in its search, however it achieves its highest performance relative to other work in the unconstrained setting.

Acknowledgments

This material is based upon work supported by the National Geospatial-Intelligence Agency under Contract No. HM0476-21-C-0004. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NGA, DoD, or the US government. Approved for public release, NGA-U-2025-00132.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 2

[2] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth. *arXiv preprint arXiv:2208.02541*, 2022. 2

[3] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8585–8594, 2022. 2

[4] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21621–21631, 2023. 2

[6] Andreas Geiger, Philip Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. In *Robotics: Science and Systems Conference (RSS)*, 2013. 4

[7] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in neural information processing systems*, 34:15908–15919, 2021. 2

[8] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 2

[9] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7779–7788, 2020. 2

[10] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018. 2

[11] Mahmoud Khalil, Ahmad Khalil, and Alioune Ngom. A comprehensive study of vision transformers in image classification tasks. *arXiv preprint arXiv:2312.01232*, 2023. 5

[12] Christopher Klammer and Michael Kaess. Bevloc: Cross-view localization and matching via birds-eye-view synthesis. *arXiv preprint arXiv:2410.06410*, 2024. 2

[13] Ted Lentsch, Zimin Xia, Holger Caesar, and Julian F. P. Kooij. SliceMatch: geometry-guided aggregation for cross-view pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5, 7

[14] Songlian Li, Zhigang Tu, Yujin Chen, and Tan Yu. Multi-scale attention encoder for street-to-aerial image geolocalization. *CAAI Transactions on Intelligence Technology*, 8(1):166–176, 2023. 2

[15] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 2

[16] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5007–5015, 2015. 2

[17] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 2

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 3, 6

[19] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 2

[20] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3163–3172, 2021. 2

[21] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019. 2

[22] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. OrienterNet: visual localization in 2d public maps with neural matching. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[23] Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[24] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[25] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[26] Yujiao Shi, Xin Yu, Liu Liu, Dylan Campbell, Piotr Koniusz, and Hongdong Li. Accurate 3-dof camera geo-localization via ground-to-satellite image matching. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):2682–2697, 2022. 2

[27] Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-dof ground-to-satellite camera

localization accuracy via geometry-guided cross-view transformer. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 5, 6, 7

[28] Hui Shuai, Lele Wu, and Qingshan Liu. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4122–4135, 2022. 2

[29] Agarwal Siddharth, Vora Ankit, Pandey Gaurav, Williams Wayne, Kourous Helen, and McBride James. Ford Multi-AV seasonal dataset. *arXiv*, 2003.07969, 2020. 4

[30] Zhenbo Song, Jianfeng Lu, Yujiao Shi, et al. Learning dense flow field for highly-accurate cross-view camera localization. *Advances in Neural Information Processing Systems*, 36:70612–70625, 2023. 5, 6

[31] Yong Tang, Qiang Huang, and Yingying Zhu. C2f-ccpe: Coarse-to-fine cross-view camera pose estimation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 5, 6, 7

[32] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3616, 2017. 2

[33] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. 2

[34] Aäron van den Oord, Yazhe Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv*, b227f3e4c0dc96e5ac5426b85485a70f2175a205, 2018. 4

[35] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[36] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5722–5731, 2021. 2

[37] Shan Wang, Yanhao Zhang, Akhil Perincherry, Ankit Vora, and Hongdong Li. View consistent purification for accurate cross-view localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[38] Shan Wang, Chuong Nguyen, Jiawei Liu, Yanhao Zhang, Sundaram Muthu, Fahira Afzal Maken, Kaihao Zhang, and Hongdong Li. View from above: Orthogonal-view aware cross-view localization. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 5

[39] Xiaofeng Wang, Zheng Zhu, Guan Huang, Fangbo Qin, Yun Ye, Yijia He, Xu Chi, and Xingang Wang. Mvster: Epipolar transformer for efficient multi-view stereo. In *European Conference on Computer Vision*, pages 573–591. Springer, 2022. 2

[40] Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 5

[41] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 70–78, 2015. 2

[42] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–9, 2015. Acceptance rate: 30.3%. 1, 2, 3, 4, 6, 7

[43] Zimin Xia and Alexandre Alahi. Fgˆ 2: Fine-grained cross-view localization by fine-grained feature matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6362–6372, 2025. 5, 6, 7

[44] Zimin Xia, Olaf Booij, Marco Manfredi, and Julian F. P. Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 5, 6, 7

[45] Zimin Xia, Olaf Booij, and Julian F. P. Kooij. Convolutional cross-view pose estimation. *arXiv*, 2303.05915, 2023. 1, 2, 5, 6, 7

[46] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3333–3343, 2022. 2

[47] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021. 2

[48] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024. 2

[49] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34: 13153–13164, 2021. 2

[50] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13760–13769, 2022. 2

[51] Jie Zhu, Bo Peng, Wanqing Li, Haifeng Shen, Zhe Zhang, and Jianjun Lei. Multi-view stereo with transformer. *arXiv preprint arXiv:2112.00336*, 2021. 2

[52] Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: cross-view image geo-localization beyond one-to-one retrieval. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4, 5, 6, 7

[53] Sijie Zhu, Mubarak Shah, and Chen Chen. 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)*, pages 1162–1171, 2022. 2