Crossview Registered Multiview Pose Estimation

Alexander Wollam Dev Gupta Nathan Jacobs
Washington University in St. Louis

Figure 1. Most camera pose estimation work either specifically looks at cross-view pose estimation with a single aerial/ground image pair or multi-view pose estimation with imagery of all the same form. In this work, we explore extending to the joint setting of multi-view ground imagery within a reference aerial image.

Abstract

Cross-view pose estimation predicts the 3 Degrees-of-Freedom (3DoF) pose, including yaw, ground-plane translation, and scale, of ground images within an aerial bird's-eye reference frame. While prior work typically uses a single ground-view image, practical settings like robotics, autonomous driving, and urban mapping often provide multiple nearby images with shared spatial structure. We leverage recent advances in multi-view pose estimation and 3D reconstruction to jointly estimate poses for multiple ground images in parallel. We adapt the VGGT multi-view 3D reconstruction framework and introduce a dataset aligning Mapillary ground-level imagery with NAIP aerial imagery for large-scale supervision. This work evaluates how the multi-view formulation transfers to cross-view pose estimation and compares it against single-view methods.

1. Introduction

Cross-view pose estimation lies at the intersection of cross-view image geo-localization and fine-grained camera pose recovery. While traditional cross-view geo-localization focuses primarily on determining where a ground image was captured by matching it to an aerial database, recent work has shown the importance of additionally reasoning about orientation for downstream tasks [32]. This shift has led to a more refined problem setting in which the goal is to recover the 3DoF pose of a ground-level image within an aerial patch, motivating applications in autonomous driving, robotics, augmented reality, and fine-grained mapping. Existing methods typically rely on feature descriptors [3, 11, 16, 30] or projections into satellite space [2, 6, 15, 17, 18, 22, 24, 25, 28, 29] to perform matching-based optimization, often assuming constrained intrinsics, fixed alignment, or homogeneous environments.

In practice, however, many cross-view applications nat-

urally provide multiple ground-level images in close proximity. For example, autonomous vehicle sequences, street-level mapping services, and crowd-sourced image collections routinely capture overlapping views of the same general location. Yet most cross-view pose estimation approaches operate strictly on single ground images, leaving substantial geometric information unexploited. Multi-ground-view settings present an opportunity to enforce geometric consistency, reduce ambiguity, and recover richer and more reliable pose estimates—particularly in challenging regions where appearance gaps or perspective distortions hinder single-view matching.

Concurrently, multi-view camera pose estimation has undergone rapid progress driven by modern learned 3D reconstruction approaches. Early methods such as DUSt3R [23] introduced feed-forward pointmap regression to recover geometry and relative pose directly from image pairs, reducing dependence on keypoint pipelines. Follow-up work either improved geometric grounding through learned matching [7] or addressed scaling and global consistency across many views [1, 31]. More recent transformer-based models such as VGGT [21], MapAnything [4], and OmniVGGT [12] demonstrated the ability to regress intrinsics, extrinsics, depth, and point tracks jointly from variable-size image sets, offering a flexible and powerful foundation for multi-view pose inference. These advances suggest a promising path for leveraging multi-image context within cross-view settings.

In this paper, we propose an extension of cross-view pose estimation that integrates these modern multi-view methods within the cross-view framework. Rather than predicting only a single ground-view 3DoF pose, we reformulate the problem to recover the poses poses for a set of nearby ground images with respect to a single aerial reference. To support this, we construct a new public dataset consisting of Mapillary ground imagery registered to NAIP aerial images, explicitly designed to evaluate multi-ground-view cross-view alignment. Furthermore, we introduce a memory-efficient extension of VGGT that incorporates a dedicated aerial-branch encoder to jointly localize and orient multiple ground images relative to an overhead view.

Our contributions are threefold:

- We create the first public dataset enabling cross-view registration in a multi-ground-view setting.
- We propose a novel approach that leverages and extends recent multi-view reconstruction methods for cross-view pose estimation.
- We explore the effectiveness of our method on our introduced dataset.

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 [9, 10, 19, 26, 27, 32]. 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 [3, 11, 16, 30]. While initial attempts at this task [9, 27] 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 [16], and this has been combined with the use of orientation considerations and newer deep architectures [30, 33]. Additionally, synthesis from one view to the other has also been explored as a way to close these gaps [8, 13, 17, 20]. 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 [32] 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 [32] introducing the VIGOR benchmark. Most approaches leverage projections into satellite space from which matching-based optimization can take place [2, 6, 15, 17, 18, 22, 24, 25, 28, 29], 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 [2, 14, 18]. 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 [5]. A downside of these BEV-based approaches, however, is that performing a dense search of candidate poses becomes expensive. Other approaches accelerate this [6, 25, 29] by aggregating features into a single vector [29], pre-computing masks [6], and reformulating the matching procedure as homography estimation [25]. Alternatively, keypoints have also been used to refine pose matching by detecting and projecting them into the satellite space [22, 24], 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.

Multi-View camera pose estimation: Recent progress in learned 3D reconstruction has led to a new class of approaches that jointly estimate multi-view geometry and camera pose without relying solely on traditional

	Vo [?]	CVACT [?]	CVUSA [?]	VIGOR [?]	(proposed)
Satellite images	$\sim 450,000$	128,334	44,416	90,618	$\sim 100,000$
Panoramas in total	$\sim 450,000$	128,334	44,416	238,696	$\sim 500,000$
Panoramas after balancing	-	-	-	105,214	-
Street-view GPS locations	Aligned	Aligned	Aligned	Arbitrary	Arbritrary
Full panorama	No	Yes	Yes	Yes	Yes
Multiple cities	Yes	No	Yes	Yes	Yes
Orientation information	Yes	Yes	Yes	Yes	Yes
Evaluation in terms of meters	No	No	No	Yes	Yes
Seamless coverage on area of interest	No	No	No	Yes	Yes
Number of references covering each query	1	1	1	4	4

Table 1. Comparison of various features between our proposed and existing datasets

optimization-heavy SfM pipelines. Early work in this direction is represented by DUSt3R [23], which introduced a pairwise pointmap regression framework capable of producing depth, correspondences, and relative pose directly from image pairs. This formulation demonstrated that pose estimation could be treated as a feed-forward geometric prediction problem, reducing reliance on extensive keypoint matching and bundle adjustment. Subsequent approaches have built upon this idea either by improving the geometric grounding of learned correspondences or by scaling DUSt3R-style reasoning to many views. MASt3R [7] strengthens the matching component by explicitly tying learned correspondences to metric 3D structure, resulting in more reliable pose estimation for localization tasks. Other work tackles scalability: Fast3R [31] processes large image collections in a single forward pass to avoid quadratic pairwise computations, while MUSt3R [1] extends the DUSt3R design with efficient global alignment strategies for consistent multi-view reconstruction. A parallel direction focuses on unifying multi-view prediction into a single feed-forward model. VGGT [21] directly regresses intrinsics, extrinsics, depth, and point tracks from one or many views, demonstrating that camera pose can be inferred as part of a broader geometric prediction suite. This unification has led to more general systems, such as MapAnything [4], which supports multiple 3D tasks under a single architecture, and OmniVGGT [12], which extends these ideas to multi-modal inputs and shows improved robustness across a range of multi-view geometry settings.

3. Dataset

Problem Statement. Given a ground-level (GL) query image inside an area-of-interest (AOI), we wish to geolocate and orient that image at a high precision level. At our means are additional high-resolution GL imagery geolocated within the AOI and a birds-eye-view (BEV) satellite image of the AOI with each containing rich metadata. As such, we wish to perform simultaneous 3 degrees-of-freedom (DoF) cross- and multi-view pose estimation, an extension of two previously disjoint tasks.

Motivation. Vast amounts of current remote sensing data is available and used for various geo-related tasks. These satellite imagery are geo-tagged with metadata relevant to that image. Utilizing this easily accessible, vast, consistent, and high-resolution imagery is essential to match existing query images to their accurate pose. Similarly, high-quality imagery is also available for GL images. Unfortunately, this problem statement benefits from rich data to supplement accurate pose estimation, and previous papers fail to address this. To solve this issue requires combining the two major sources: both cross and multi views. We wish to construct a dataset for this new format that not only allows for these two major sources, but provides it open-source at scale.

Dataset Curation. To construct this large-scale dataset, we aggregate data from various diverse urban, suburban, and rural areas across different cities/regions to ensure geographical and architectural diversity. For the ground-level (GL) multi-view component, we source sequences of high-resolution street-level imagery, ensuring spatial proximity and visual overlap between neighboring frames. Each GL image is meticulously paired with its corresponding ground-truth 3-DoF pose (latitude, longitude, and azimuth/heading) and camera intrinsics. For the cross-view component, we acquire high-resolution, orthorectified birds-eye-view (BEV) satellite imagery corresponding to the exact area-of-interest (AOI) bounding boxes of the GL networks. To guarantee data quality, we implement an automated filtering pipeline to remove occluded scenes (e.g., tunnels, dense tree canopies), indoor images, and misaligned GPS tags, followed by rigorous manual verification. The rich metadata provided includes time-of-day, weather conditions, lighting priors, and geometric camera parameters. In total, the dataset comprises 500,000+ GL images and 100,000+ BEV aerial maps, partitioned into strictly disjoint training, validation, and testing geographic splits to properly evaluate model generalization capability.

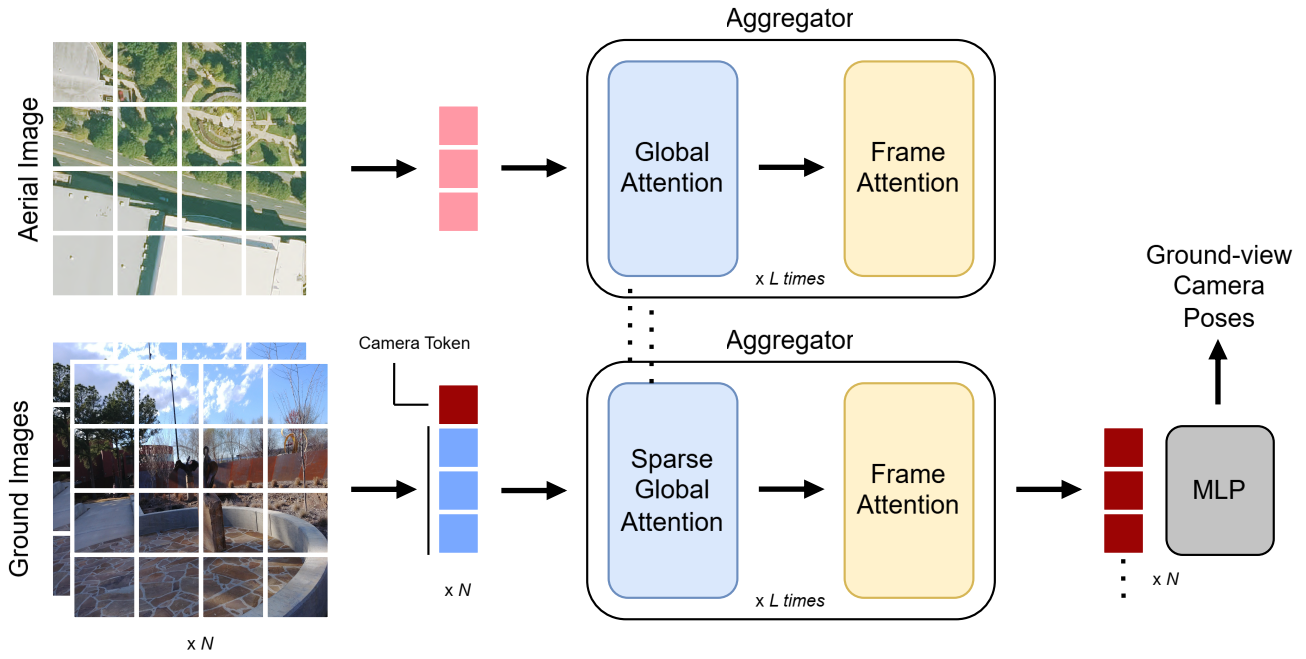

Figure 2. An overview of the VGGT model with our modifications.

Extensions. While the primary objective of this dataset is simultaneous cross- and multi-view 3-DoF pose estimation, its scale, rich metadata, and unique structural design readily support a variety of secondary tasks and future research directions. The inclusion of multi-view GL imagery alongside aligned BEV satellite data provides an ideal testbed for cross-view novel view synthesis (e.g., cross-view NeRFs or 3D Gaussian Splatting) and large-scale 3D scene reconstruction. Additionally, the dataset can be utilized for image-based semantic layout estimation, where models infer top-down maps directly from GL panoramas. Because we include environmental metadata, researchers can also evaluate the robustness of visual geo-localization algorithms against seasonal changes, variable illumination, and transient occlusions. Finally, the open-source nature of this dataset allows for seamless future integration with supplementary modalities, such as digital elevation models (DEMs) or temporal satellite sequences, paving the way for 6-DoF pose estimation and spatio-temporal geo-localization.

4. Method

Our proposed approach enables performing cross-view pose estimation in a multi-ground-view setting. It does this by augmenting VGGT to use an aerial image through an explicit aerial reference branch, and modifying the global attention layers to sparsely attend across ground views (see figure 2). Our method maintains high performance while also minimizing memory costs.

4.1. Aerial Branch

In the original multi-view stereo problem setting of VGGT, it predicts the relative camera pose and 3D point maps with respect to its first image. In contrast, in the cross-view pose estimation setting, rather than predict the relative pose to a particular image we instead define a global coordinate space of the aerial image, from which ground image poses are predicted.

To account for this change from predicting a 6-DoF relative camera pose to a 3-DoF absolute pose within the aerial image, with the VGGT architecture, we create a separate branch for the aerial image that maintains the previous VGGT structure, but with independent weights. Specifically, we clone the initial weights and setup to the aerial branch, then modify the global attention layers to only take the aerial patches as queries. This branch separation is done to enable optimization of these weights for this new setting while maintaining learned behaviour by selectively freezing existing weights for the ground images.

4.2. Sparse global attention

One issue with the existing VGGT architecture is that its use of global attention layers incurs memory costs that scale quadratically with the number of images input to the model. To get around this, we propose leveraging the cross-attention maps between the aerial and ground views to inform likely similar patches between ground views that can

then selectively attend to one another sparsely (see figure 3).

To do this, we propose replacing each global attention layer in the ground branch with a sequential cross-aerial attention and sparse global-attention layer. Here, the cross-aerial attention layer first performs cross attention between the ground and aerial views. Then, the resulting attention map is utilized to sparsify the global attention as follows.

First, the attention map is aggregated across heads. Second, for each aerial patch, we select the top k_f ground view frames based on their most attending patch. Then, for each selected frame per aerial patch, we select the top k_{gp} ground image patches to get $k_f * k_{gp}$ ground image patches per aerial patch. Finally, for each ground patch, we select the top k_{ap} aerial patches associated with it. This results in a mapping of each ground image patch to an associated set of $k = k_f * k_{gp} * k_{ap}$ total ground patches from which the sparse global attention layer then attends.

By sparsifying which patches each ground patch can attend to just this selected k, we can leverage the aerial patches for an intermediate 'binning' of ground patches to significantly reduce the memory overhead.

4.3. Training

To train our model, we first initialize shared weights from the original VGGT model to leverage it's pre-learned behaviour. This is done since our problem formulation only looks at pose estimation, which provides far less supervision than the full set of predictions the base model utilizes. For the 3-DoF pose prediction, we augment VGGT's 6-DoF prediction head by simply taking that output and transforming it to the associated 3-DoF format.

To maintain performance from the base model, we only optimize the aerial-branch's weights as well as the ground branch's sparse global attention layers, and freeze all of the rest.

5. Experiments

In this section, we introduce the datasets used for evaluation, our implementation details, and the evaluation metrics . Then, we perform an ablation over different parameters of the model.

5.1. Datasets

For our experiments, we train and evaluate our model over both the grid and non-grid versions of our dataset. We then perform ablations over the grid version.

5.2. Implementation Details

For implementing our model, we follow the strategy discussed in section 4.3 and first initialize the model off of the VGGT-1B checkpoint provided by VGGT [21]. We then duplicate the Aggregator weights to form the separate aerial and ground view branches. Following, we replace the ground branch's global attention layer with our sparse global attention layer whose weights are similarly initialized as the original global layer's where possible.

For selecting the number of ground images used, we sample 6 ground images with replacement for each sample. We choose to use replacement to account for the fact that a subset of samples have less than 6 images, so duplication of those images are necessary to maintain a consistent shape across samples.

For parameterizing our top-K selection, we choose $K <= n$ where n is the number of patches in a single ground-view image; this is done to enforce a memory complexity of the sparse global attention layer to be in $O(f*n^2)$, where f is the number of ground view frames—this way it has a similar memory footprint as the frame self-attention layers. For our data, we resize images to 504x504 pixels, leading to 36x36 patches, for $n = 1,296$ patches per image total. For simplicity, we choose a maximum of $k = 1,000$ for the main evaluations, and ablate over less. We then choose a balanced approach between k_f, k_{gp}, k_{ap} of $k = k_f * k_{gp} * k_{ap}$, and set all to 10.

5.3. Evaluation Metrics

In the evaluations below, we use the mean localization in meters and the mean orientation in degrees to analyze our model's performance.

5.4. Ablations

In this section we evaluate our model over different settings. First we evaluate over the different versions of the dataset, then we explore the impact on varying the number of ground view frames, and finally we analyze the impact of top-k selection parameters for the sparse global attention module.

Table 2. Mean Localization and Orientation Errors

Dataset	Loc. Error (m)	Ori. Error (°)
Grid Dataset	7.32	4.58
Non-Gridded Dataset	7.01	4.46

5.4.1. Dataset Performance

We view the performance of our model when trained and evaluated use the gridded and non-gridded versions of our dataset in table 2. Here, we see slightly improved performance with the non-gridded dataset. This likely is a result of the 'built-in' data-augmentation-like aerial shifting resulting in slightly more robust model performance.

5.4.2. Varying number of ground images

We explore the impact of varying the number of ground-view frames within the satellite image provided to the

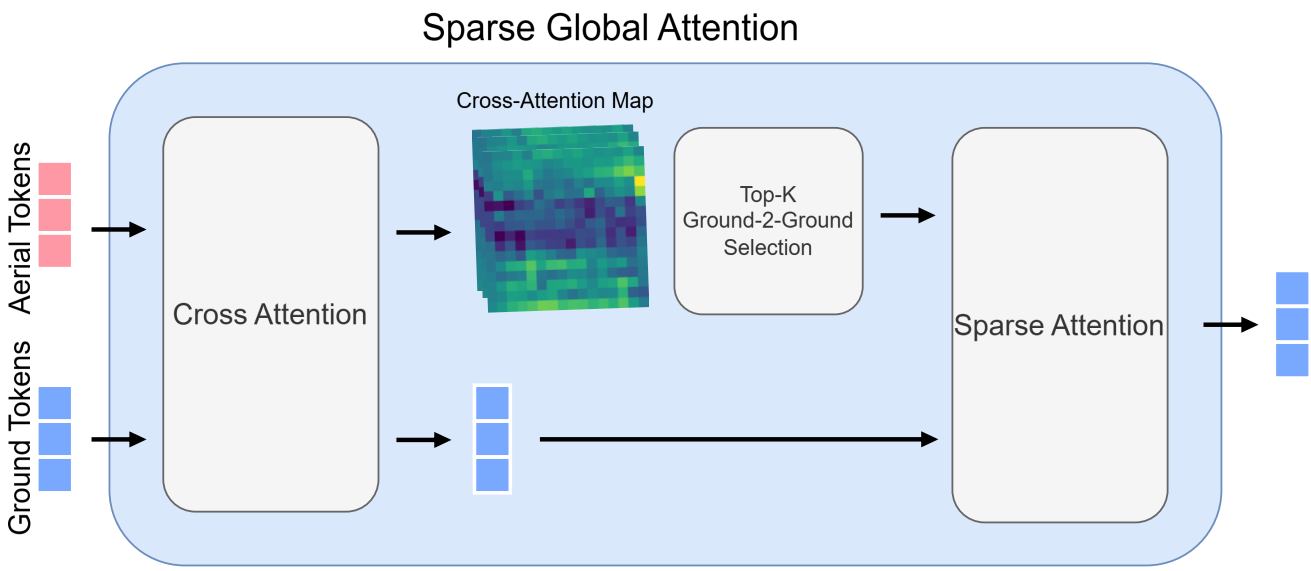

Figure 3. The structure of the sparse global attention layer. Note that first cross-attention is performed between the ground and aerial views, this the resulting attention map is used to derive a top K ground-to-ground mapping, which is used for the keys/values for the sparse global attention.

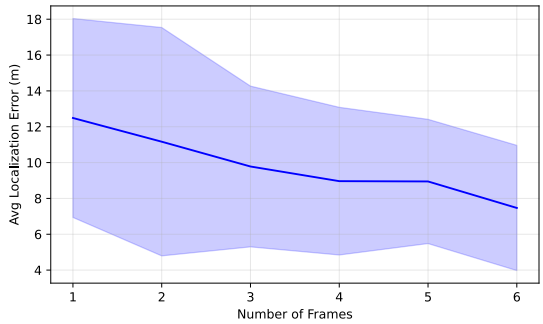

Figure 4. An overview of how the number of frames selected impact model localization performance.

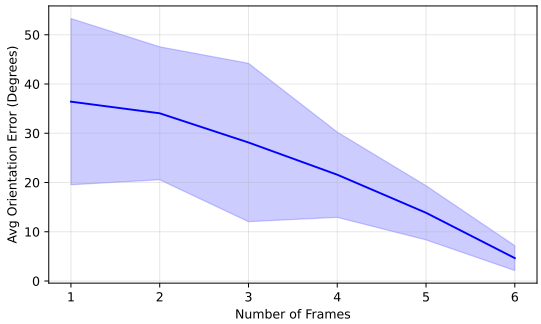

Figure 5. An overview of how the number of frames selected impact model orientation performance.

model. For localization, we can see in figure 4 that localization performance decreases as we reduce the number of frames provided. Similarly for orientation, we can see in figure 5 that orientation performance also decreases with reduced frame count. This makes sense, as the model can leverage having multiple ground views to help constrain predictions, particularly when there is ambiguity (e.g. few localizable features such as if the image is on a highway with only trees around). We also visualize the variance in prediction at the different frame counts for both figures as the shaded region around the plotted lines. Here, was can also see that the variance increases as the number of ground views decreases, which matches our expectation of fewer views being less constrained.

5.4.3. Varying top-K selection size

We also explore the impact of varying the top-k selection size. In particular, we look at the impact of the size of K. For localization, we see in figure 6 that K size has a significant impact on performance, where too small of a K results in poorer performance. We see similarly with orientation performance in figure 7. This suggests that our approximate top-K selection algorithm does not consistently select all of the important patches for the sparse global attention, resulting in less consistent performance at lower values of K. We can, however, note that the performance drop is similar to what is seen when decreasing the number of frames. We hypothesis that this is because decreasing K reduces global context for the sparse global attention in a similar manner to

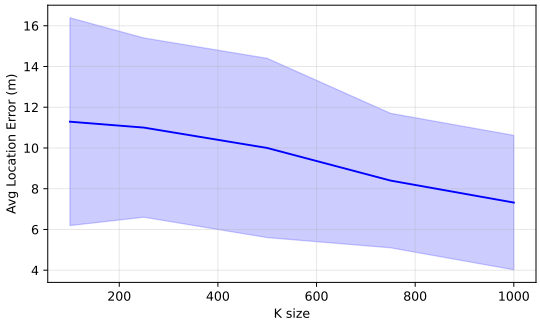

Figure 6. An overview of how the size of K selection size in sparse global attention impacts model localization performance.

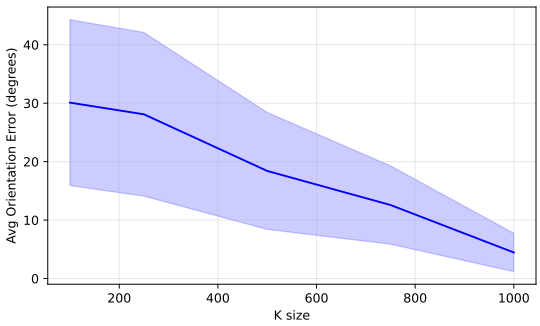

Figure 7. An overview of how the size of K selection size in sparse global attention impacts model orientation performance.

reducing the number of frames does—where selecting only one ground-view frame has a similar effect to reducing K close to zero of eliminating that context.

6. Conclusion

In this paper we introduced the new problem formulation of camera pose estimation featuring multiple ground images registered to a single aerial view. To this end, we create a new dataset using Mapillary and NAIP imagery and adapt the recent multi-view reconstruction method VGGT to this new setting. From this, we explore the performance of our model and dataset under different settings and demonstrate results that suggest this direction of research is promising.

References

[1] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1050–1060, 2025. 2, 3

[2] 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. 1, 2

[3] 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. 1, 2

[4] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 2, 3

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

[6] 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. 1, 2

[7] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2, 3

[8] 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

[9] 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

[10] 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

[11] 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. 1, 2

[12] Haosong Peng, Hao Li, Yalun Dai, Yushi Lan, Yihang Luo, Tianyu Qi, Zhengshen Zhang, Yufeng Zhan, Junfei Zhang, Wenchao Xu, et al. Omnivggt: Omni-modality driven visual geometry grounded. *arXiv preprint arXiv:2511.10560*, 2025. 2, 3

[13] 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

[14] 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

[15] 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

[16] 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. 1, 2

[17] 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. 1, 2

[18] 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

[19] 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

[20] 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

[21] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 3, 5

[22] 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

[23] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2, 3

[24] 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. 1, 2

[25] 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

[26] 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

[27] 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%. 2

[28] 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. 1, 2

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

[30] 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. 1, 2

[31] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025. 2, 3

[32] 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. 1, 2

[33] 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