BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

# Camera Orientation Estimation using Vision Transformers

Martin Kubička*

**Abstract**

Many times, we look at a photo and ask: Where was it taken? In computer vision, this problem is called geolocalization, which includes camera orientation estimation. This work focuses on camera orientation estimation, a fundamental problem in computer vision with applications in augmented reality, robotics, and autonomous driving. We estimate camera orientation (pitch, yaw and roll angles) by matching a real query image to a synthetic digital elevation model (DEM), focusing on challenging mountain environments with possible strong seasonal and weather variations. Unlike existing state-of-the-art (SOTA) methods, our transformer-based approach does not require field-of-view (FOV) information, making it applicable to common internet images. Despite this, it achieves better results. In addition, this thesis provides a comprehensive survey of geolocalization, pose estimation and orientation estimation methods, and outlines future directions including cross-domain applications, explainable approaches, and extension to full geolocalization.

*xkubic45@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

## 1. Introduction

This work focuses on camera orientation estimation, a fundamental problem in computer vision. The task is to align a synthetic DEM, given as a 360° panorama, with a real query image. The result is described by three angles: pitch, yaw, and roll. This problem has applications in AR, robotics, security, autonomous driving, and related areas [1]. It is also challenging due to the cross-domain inputs, high resolution, long training time, and limited data. So, this work can inspire solutions for similar problems. Previous works such as [1, 2, 3, 4, 5, 6, 7, 8] often use edges and other features like semantic cues, but these are not always clear, because of weather conditions and seasons. Other works [9, 10] avoid direct matching and compare learned features instead. The SOTA orientation estimation methods [1, 2] also use FOV, which is usually not available. We solve the problem using a transformer-based model without FOV, letting the network learn the best way to estimate orientation. Understanding this learned behavior is an important direction for future work. Despite the challenges, our method achieves results better than SOTA methods.

## 2. Camera Orientation Estimation

The problem solved in this work is called camera orientation estimation. As shown in Figure 1, there are two inputs: a synthetic terrain model (DEM) as a 360° panorama, and a real query image. The goal is to find the orientation of the query image within the panorama, i.e., where it is located. The orientation is described by three angles: pitch, yaw, and roll. Pitch ($\alpha$) is rotation around the X axis in the range [-90°, 90°], with 0 in the center. Yaw ($\beta$) is rotation around the Y axis in the range [-180°, 180°], with 0 in the center. Roll ($\gamma$) is rotation around the Z axis in the range [-180°, 180°], where 0 means the image is aligned with the horizon.

### 2.1 Challenges

Solving our problem involves several important challenges. First, we do not use the FOV as input, which is often unknown for real-world images. In contrast, methods like [1, 2] rely on this information. Another challenge is limited training data. We use GeoPose3K [11] and LandscapeAR [9], where the LandscapeAR required manual filtering due to incorrect samples. After cleaning, we have 17,017 panorama-image pairs. High input resolution is also crucial, as orientation estimation depends on fine details. However, this leads to slow training (several days per epoch on a single GPU), slower inference, and high memory usage, so the problem itself is complex. Besides that, we use two inputs, which increases computation time and involves a cross-domain setup (matching synthetic terrain models to real images). Although there are some neural

network approaches [12, 13], most related tasks still rely on traditional methods [1, 8, 14], making this problem less explored. Finally, an input is a 360° equirectangular panorama, which introduces distortions that can cause problems [15]. Ideally, spherical approaches [16, 17, 18] would better handle this, but they are not used in our solution.

## 3. Solution

### 3.1 Architecture

As mentioned, we solve the problem using transformer-based neural networks. The best architecture from ablation experiments is shown in Figure 4. It is partly inspired by the cross-attention approach of [19] and the classification-based approach of [20]. The model takes a panorama of size $4096 \times 2048$px and a query image of size $512 \times 512$px. Both inputs use the same pretrained PE-Spatial-Tiny [21] Vision Encoder, which is fine-tuned. Because of this, the panorama is first split into $512 \times 512$px tiles. Position encoding is then added to the tiles, and both inputs use 1D sinusoidal positional encoding. Next, a shared Cross-Attention layer is applied to learn relationships between the inputs. Then, a shared Attention Pooling layer highlights important parts and converts the features to the correct shape for prediction. The model predicts three angles: pitch, yaw, and roll, each with its own Multi-Layer Perceptron (MLP) head. The model outputs 180 classes for pitch and 360 classes for yaw and roll (one per degree). We found out that classification works better than regression for this task. Training uses Circular Huber Loss, which considers angle wrap-around (e.g., $-179°$ is $2°$ from $+179°$, not $358°$). Training progress (loss and orientation error over 34 epochs) is shown in Figure 2.

### 3.2 Experiments

As part of the solution, we present several experiments: Ablation Study, Second-Stage Refinement Model, Impact of Input Resolution, Region Bias Analysis, Impact of Data Augmentation, Model Interpretation (Failure Case Analysis and Attention Map Analysis).

## 4. Results

For comparison and evaluation of results with the method [1], we use the following orientation estimation error formula:

$$e(\mathbf{R}_{gt}, \mathbf{R}c) = \arccos\left(\frac{\mathrm{tr}\left(\mathbf{R}gt^{T}\mathbf{R}_{c}\right)-1}{2}\right)$$

where $\mathbf{R}_{gt}$ is the ground truth camera rotation matrix and $\mathbf{R}_{c}$ is the estimated rotation matrix, where the rotation matrix includes all three angles – pitch, yaw, and roll. This metric calculates the magnitude of the smallest rotation between the ground truth and the estimated rotation. The results are shown in Figure 3. The $X$ axis shows the orientation error, and the $Y$ axis shows the percentage of images within a given error. For example, our method (blue curve) reaches about 75% of images with an error up to $20°$. The evaluation is done on the GeoPose3K [11] test split. The orange curve shows results of the SOTA method [1], and the green curve shows random estimation. The graph also includes Area Under Curve (AUC). Our method achieves higher AUC (0.89 vs. 0.78) but lower precision at errors below 10°. The FOV-free design justifies this trade-off for real-world internet images where FOV is unavailable. Ideally, the curve should rise faster at small errors. This could be improved in future work, for example by using a two-step approach with initial prediction and later refinement. Our method has an average orientation error of $19.7°$ on the test set, but unfortunately this information is not available for [1]. We could not compare with other methods such as [2], because results are either not on GeoPose3K dataset or other test datasets are not available and code is also not publicly available.

## 5. Conclusions

In this work, we studied camera orientation estimation, where the goal is to align a real image with a synthetic 360° terrain model (DEM). The orientation is given by three angles: pitch, yaw, and roll. We designed our own transformer-based model, which achieved better results than SOTA methods. The main advantage is that, unlike other methods, our approach does not use FOV, which is often not available in real photos. We also provide a detailed survey of methods for geolocalization, camera pose, and orientation estimation. In addition, we manually cleaned the LandscapeAR [9] dataset by removing incorrect panorama–image pairs. We performed many experiments, including ablation studies, model interpretation, among others. Our approach can inspire solutions for similar cross-domain problems and handling combination of challenges such as high resolution. The method can be improved with architectural changes and multi-step refinement for better accuracy. It can also be extended to full geolocalization, for example, by using multimodal large language models (MLLMs) for spatially relational scene descriptions, followed by further processing with another network such as graph neural networks.

## Acknowledgements

I would like to thank my supervisor, Professor Martin Čadík, and my family and friends for their support and encouragement throughout this work.

## References


[1] Jan Brejcha and Martin Čadík. Camera orientation estimation in natural scenes using semantic cues. In *2018 3DV*, pages 208–217, 2018.

[2] Lionel Baboud, Martin Čadík, Elmar Eisemann, and Hans-Peter Seidel. Automatic photo-to-terrain alignment for the annotation of mountain pictures. In *2011 CVPR*, pages 41–48, 2011.

[3] Lorenzo Porzi, Samuel Rota Bulò, Oswald Lanz, Paolo Valigi, and Elisa Ricci. An automatic image-to-dem alignment approach for annotating mountains pictures on a smartphone. *Mach. Vis. Appl.*, 28(5):809–821, 2017.

[4] Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys. Leveraging topographic maps for image to terrain alignment. In *2012 3DIMPVT*, pages 487–492, 2012.

[5] Prospero Naval, Masayuki Mukunoki, Michihiko Minoh, and Katsuo Ikeda. Estimating camera position and orientation from geographical map and mountain image. In *38th Research Meeting of the Pattern Sensing Group, Society of Instrument and Control Engineers*, pages 9–16, 1997.

[6] Prospero Naval. Camera pose estimation by alignment from a single mountain image. *International Symposium on Intelligent Robotic Systems*, pages 157–163, 2010.

[7] Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *2012 ECCV*, pages 517–530, 2012.

[8] Yi Chen, Gang Qian, Kiran Gunda, Himaanshu Gupta, and Khurram Shafique. Camera geolocation from mountain images. In *2015 Fusion*, pages 1587–1596, 2015.

[9] Jan Brejcha, Michal Lukáč, Yannick Hold-Geoffroy, Oliver Wang, and Martin Čadík. Landscapear: Large scale outdoor augmented reality by matching photographs with terrain models using learned descriptors. In *2020 ECCV*, volume 12374 of *Lecture Notes in Computer Science*, pages 295–312. 2020.

[10] Jan Tomešek, Martin Čadík, and Jan Brejcha. CrossLocate: Cross-modal Large-scale Visual Geo-Localization in Natural Environments using Rendered Modalities. In *2022 WACV*, pages 2193–2202, 2022.

[11] Jan Brejcha and Martin Čadík. GeoPose3K: Mountain landscape dataset for camera pose estimation in outdoor environments. *Image and Vision Computing*, 66:1–14, 2017.

[12] Xingyi He, Hao Yu, Sida Peng, Dongli Tan, Zehong Shen, Hujun Bao, and Xiaowei Zhou. Matchanything: Universal cross-modality image matching with large-scale pre-training. In *Arxiv*, 2025.

[13] Nam N. Vo and James Hays. Localizing and orienting street views using overhead imagery. In *2016 ECCV*, pages 494–509, 2016.

[14] Hang Chu, Andrew Gallagher, and Tsuhan Chen. Gps refinement and camera orientation estimation from a single image and a 2d map. In *2014 CVPRW*, pages 171–178, 2014.

[15] W. Yang, Y. Qian, J. Kämäräinen, F. Cricri, and L. Fan. Object detection in equirectangular panorama. In *2018 ICPR*, pages 2190–2195, 2018.

[16] Oscar Carlsson, Jan E. Gerken, Hampus Linander, Heiner Spieß, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. HEAL-SWIN: A Vision Transformer On The Sphere. In *2023 CVPR*, pages 6067–6077, 2023.

[17] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. In *International Conference on Learning Representations*, 2018.

[18] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *2018 ECCV*, pages 518–533, 2018.

[19] Shay Dekel, Yosi Keller, and Martin Cadik. Estimating extreme 3d image rotations using cascaded attention. In *2024 CVPR*, pages 2588–2598, 2024.

[20] Hana Bezalel, Dotan Ankri, Ruojin Cai, and Hadar Averbach-Elor. Extreme rotation estimation in the wild. In *2025 CVPR*, pages 1061–1070, 2025.

[21] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.