VICI: VLM-Instructed Cross-view Image-localisation

Xiaohan Zhang*
University of Vermont
Burlington, USA
xiaohan.zhang@uvm.edu

Tavis Shore*
University of Surrey
Guildford, UK
t.shore@surrey.ac.uk

Chen Chen
University of Central Florida
Orlando, USA
chen.chen@crcv.ucf.edu

Oscar Mendez
Locus Robotics
Boston, USA
omendez@locusrobotics.com

Simon Hadfield
University of Surrey
Guildford, UK
s.hadfield@surrey.ac.uk

Safwan Wshah
University of Vermont
Burlington, USA
safwan.wshah@uvm.edu

Abstract

In this paper, we present a high-performing solution to the UAVM 2025 Challenge [24], which focuses on matching narrow Field-of-View (FOV) street-level images to corresponding satellite imagery using the University-1652 dataset. As panoramic Cross-View Geo-Localisation nears peak performance, it becomes increasingly important to explore more practical problem formulations. Real-world scenarios rarely offer panoramic street-level queries; instead, queries typically consist of limited-FOV images captured with unknown camera parameters. Our work prioritises discovering the highest achievable performance under these constraints, pushing the limits of existing architectures. Our method begins by retrieving candidate satellite image embeddings for a given query, followed by a re-ranking stage that selectively enhances retrieval accuracy within the top candidates. This two-stage approach enables more precise matching, even under the significant viewpoint and scale variations inherent in the task. Through experimentation, we demonstrate that our approach achieves competitive results - specifically attaining R@1 and R@10 retrieval rates of 30.21% and 63.13% respectively. This underscores the potential of optimised retrieval and re-ranking strategies in advancing practical geo-localisation performance. Code is available at github.com/tavisshore/VICI.

CCS Concepts

• **Computing methodologies** → **Vision for robotics**.

Keywords

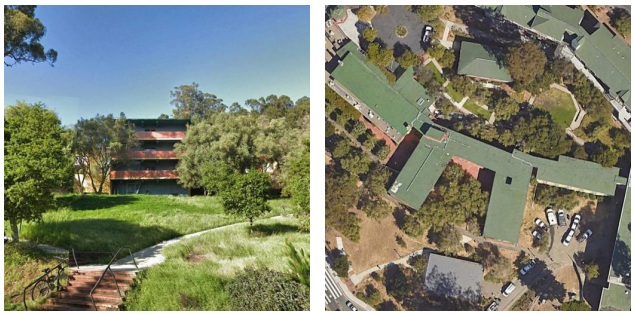
Image Localisation, Cross-View Geo-Localisation, Vision-Language Model, Image Retrieval

ACM Reference Format:
Xiaohan Zhang, Tavis Shore, Chen Chen, Oscar Mendez, Simon Hadfield, and Safwan Wshah. 2025. VICI: VLM-Instructed Cross-view Image-localisation. In *Proceedings of the 3rd International Workshop on UAVs in Multimedia: Capturing the World from a New Perspective (UAVM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3728482.3757386

*Both authors contributed equally to this research.

UAVM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-1839-7/2025/10
https://doi.org/10.1145/3728482.3757386



The ground image building's distinctive multi-level structure with red/orange balconies matches the building visible in the center-left of the satellite, which has a similar tiered appearance and dark base. The grassy slope leading up to the building, the winding path, and the surrounding trees in the ground image are consistent with the landscape features around this building in the satellite image. The ground camera was likely positioned on the stairs or the path in the lower-left portion of the satellite image, looking northeast towards the building.

Figure 1: VICI localisation example: *Top left:* query image, *Top right:* top retrieved satellite image. *Bottom:* justification for this satellite image being re-ranked to Top-1.

1 Introduction

Localisation is a fundamental requirement in mobile robotics, as agents must ascertain their position before executing assigned tasks. These generally rely on Global Navigation Satellite Systems (GNSS) for localisation; however, this approach becomes unreliable in urban canyons or conflict zones, where signal obstruction, multipath effects, or deliberate jamming degrade performance. Cross-View Geo-Localisation (CVGL) offers a robust alternative to address this issue by inferring the location of a street-level image by matching it to a corresponding geo-tagged satellite image in which it appears. In most existing works [2, 16, 32, 33, 38], limited Field-of-View (FOV) ground query images are not fully explored due to the extreme lack of contextual surrounding information. Instead, existing methods primarily focused on panoramic imagery, leveraging its wide FOV to extract descriptive features and optimise matching accuracy. However, the majority of mobile robots, from autonomous vehicles to warehouse platforms, are equipped with limited-FOV cameras, thereby impeding the practical deployment of such systems [27, 34]. But, directly applying such a feature-matching paradigm to limited-FOV images often introduces noise,

making features less distinguishable, leading to a reduction in CVGL performance [12, 17].

The recent emergence of Large Language Models (LLMs) and Vision-Language Models (VLMs), including ChatGPT [13], Gemini [20], and LLaMA [22], has showcased the strength of these foundation models in image understanding [8], visual question answering [5], and text-to-image synthesis [21]. One idea to alleviate the above-mentioned problem is to leverage the reasoning capability of such models to match the query image against the reference satellite database, providing justification for the matching and offsetting the loss of information. However, naively applying VLMs on the whole reference database is costly and inefficient. To tackle this issue, we propose VICI, **V**LM-**I**nstructed **C**ross-view **I**mage-localisation, a novel two-stage VLM-powered CVGL model. The first stage extracts visual features from both ground and satellite views - predicting a coarse ranking for the ground query. To alleviate the over-fitting issue, we incorporate drone images to augment the satellite data. In the next stage, the Top-10 retrieved candidate satellite images are re-ranked by a Vision-Language Model (VLM), which also takes the query image and a curated prompt as input. In this manner, VICI not only improves the localisation accuracy but also maintains the computational overhead at a reasonable scale. Furthermore, VICI not only re-ranks the predictions but also provides the justifications for the re-ranking decisions. One localisation example with justifications is shown in Figure 1.

We achieve very competitive performance in the challenge [24], attaining R@1 and R@10 retrieval rates of 30.21% and 63.13% respectively.

In summary, our research contributions are:

- Introduction of VICI, a novel two-stage CVGL framework that integrates VLMs to go beyond traditional feature similarity methods. Our approach not only substantially improves localisation accuracy but also introduces interpretable reasoning through language-based justifications.
- A novel data augmentation technique that incorporates high-angle drone imagery within the satellite image branch to enhance model robustness and generalisation.
- Extensive experiments demonstrate the competitive performance of our VICI on the UAVM 2025 challenge [24]. We also provide quantitative and qualitative evidence for the superiority of the novel two-stage design, illustrating a potential new research direction for the field of CVGL.

2 Related Work

Cross-View Geo-Localisation: The deep learning era of CVGL began with Workman and Jacobs [28], who demonstrated the efficacy of Convolutional Neural Networks (CNNs) for correlated feature extraction across different viewpoints. CVGL datasets primarily consisted of panoramic street-level and satellite image pairs, including CVUSA [29], CVACT [9], and VIGOR [37]. Recognising the need to better model real-world scenarios, Shi et al. [16] introduced the limited Field-of-View (FOV) crops into the CVGL research. Shore et al. [19] proposed representing data as a graph, leveraging connectivity information to enhance performance, and subsequently [18] increasing discriminability by adding reference street-level images to this representation. More recently, to improve overall generalisation and address limited dataset diversity, Huang et al. released CV-Cities [6], encompassing a wider range of global city scenes.

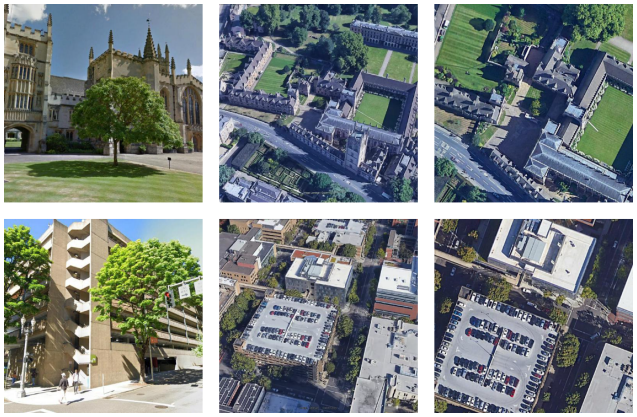


Figure 2: Street-level, drone, and satellite images from various locations, illustrating how the drone imagery provides feature continuity between the viewpoints.

Backbone feature extractors play a vital role in CVGL. Recently, transformers [23] were introduced to CVGL by two seminal works[30, 36]. Yang et al. [30] combined a ResNet backbone with a vanilla ViT encoder. Zhu et al. [36] proposed a transformer that uses an attention-guided non-uniform cropping to remove uninformative areas. Zhu et al. [38] introduced an attention-based backbone, representing long-range interactions among patches and cross-view relationships with multi-head self-attention layers. Sample4Geo [2] proposed two sampling strategies, sampling geographical sampling and hard sample mining to improve CVGL accuracy. In GeoDTR [32, 33], Zhang et al. decouple geometric information from raw features, learning spatial correlations within visual data to improve performance.

Vision-Language Models for Geo-localisation: VLMs are increasingly being used in image localisation for their logical reasoning capabilities. Initially, they were fine-tuned to operate with street-level images, combining viewed features to logically determine location. GeoReasoner [7] is a two-stage fine-tuned large VLM that mimics human reasoning from geographic clues to accurately predict locations from street-level images. Ye et al. [31] introduce a text-guided CVGL method that retrieves satellite images using natural language descriptions of street-level scenes, enabling localisation without requiring a query image. Dagda et al. propose GeoVLM [1], using a Vision-Language Model to perform zero-shot CVGL by re-ranking candidate satellite-ground image pairs using language-based scene descriptions. Whereas [1] employs VLMs to categorise specific features within images, our approach leverages VLMs to directly compare image pairs, enabling them to re-rank retrievals and provide justification for their decisions.

3 Methodology

The proposed method, VICI, operates in two stages: 1) Coarse Retrieval, extracting feature embeddings from input images - ranking reference embeddings by similarity to the query, and 2) VLM Re-ranking and Reasoning, re-ranking the candidates from step 1)

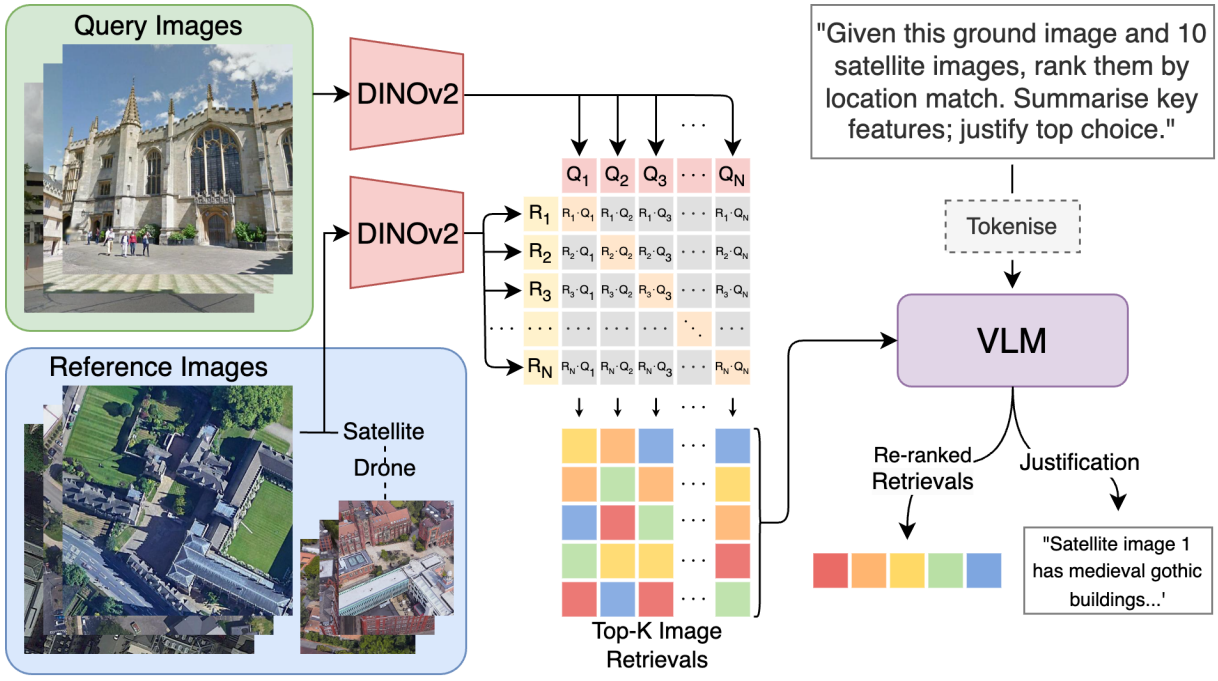


Figure 3: Overview of the architecture: Features extracted by separate DINOv2 branches, references retrieved by descending similarity, re-ranked and justified by passing through a VLM with a text prompt.

using a VLM prompted to focus on salient static features shared across viewpoints. An overview of the proposed technique is shown in Figure 3.

3.1 Stage I: Coarse Retrieval

In Stage I, we employ a Siamese network f, without weight sharing, parametrised by θ to minimise the domain gap between street-level and satellite image features, generating embeddings η_t where $t \in \{street, sat\}$. All input images I are in RGB space: $I \in \mathbb{R}^{3\times W\times H}$, and are resized to $W \times H$, *where* $H = W, W \in \{224, 384, 448, 518\}$ for different backbone extractor configurations. Mathematically, this process can be defined as follows,

$$\eta_t = f_\theta\left(I_t\right), t \in \{street, sat\} \tag{1}$$

During inference, reference embeddings are precomputed and stored offline, querying this database for retrievals during online operation.

Drone Perspective Augmentation: The challenge dataset, University-1652 [35], curates street, drone, and satellite views. Although this challenge focuses on street-to-satellite CVGL, drone imagery provides intermediate, low-altitude oblique views that are significantly more similar to street-level imagery than traditional nadir satellite views. More specifically, unlike satellites that observe scenes from near-vertical angles at high altitudes, drones capture structures and terrain from oblique angles and much lower altitudes—typically tens to hundreds of metres—making them closer in both scale and viewpoint to ground-based images (as shown in Figure 2). Thus, drone-view images can assist in bridging the domain gap between ground and satellite viewpoints.

Inspired by this, during training, we randomly feed drone-view images into the satellite branch in place of the satellite images according to a probability P, alongside their corresponding street-level images. Experimentation demonstrates how the geometric and visual similarities between drone and ground views help narrow the domain gap between aerial and ground perspectives, enabling more effective feature matching and correspondence estimation. The improved alignment in perspective mitigates issues like occlusion, foreshortening, and extreme viewpoint disparity. As a result, drone-satellite fusion enhances spatial reasoning and improves CVGL accuracy, particularly in dense urban or structurally complex scenes.

3.2 Stage II: VLM Re-ranking and Reasoning

VLMs recently demonstrated strong performance in recognising objects, interpreting scenes, and aligning visual content with natural language [8, 20]. In the second stage of VICI, we leverage this capability by feeding the top-10 retrievals from the reference database into a VLM along with a curated prompt and the query street-level image. This prompts the VLM to logically reason about the candidates, produce a more accurate ranking, and justify the decisions. Below is a simplified version of the text prompt used to re-rank the retrieved satellite images:

Given one ground image and 10 satellite images, identify which satellite image matches the ground location. Summarise the ground image and each satellite image, focusing on key features (streets, buildings, etc.). Then, compare the ground image with each satellite image as well as the summarisation. Rank these 10 satellite images by likelihood [1–10]. Justify the top choice with matching features and estimated camera position.

After receiving the response, we extract the re-ranked results from the VLM output, along with the justification for the top choice. For the full prompt, please refer to our code.

3.3 Implementation Details

Stage I: The stage I of VICI is implemented in PyTorch [15] and trained with InfoNCE loss [2] for 100 epochs and batch size of 32 using an AdamW [11] optimiser with an initial learning rate of 1e-5 and an exponential scheduler with gamma of 0.9. We employed a wide variety of backbone feature extractors such as ConvNext [10], ViT [3], and DINOv2 [14]. Training and testing of Stage I are conducted on 4 AMD MI210 accelerators.

Stage II: The second stage of VICI for re-ranking and reasoning is performed with Google's Gemini 2.5 [4]. We chose two model variants, Gemini 2.5 Flash and Gemini 2.5 Flash Lite, to balance efficiency and cost. We set the temperature to 0 to have a fixed output and better reproducibility. We leverage the structured output functionally [1] to make the output follow a JSON structure. For more details, please refer to the corresponding code.

4 Evaluation

Dataset: This challenge [24] utilises the University-1652 dataset [35] for benchmarking purposes. This dataset contains image sets of 1,652 unique university buildings: 701 for training and the rest for testing. Each image set contains a single satellite image featuring the building, 54 drone images captured with an ascending circling trajectory, and a few street-level, limited-FOV images cropped from street view panoramas. All images share the same resolution of 512×512. To have a fair comparison and illustrate the power of the proposed VICI under the case of limited training data, we did **NOT** include extra training data, although it is allowed in this challenge.

Backbone Comparison: The first experiment aims to identify the most effective feature extraction architecture for this challenging limited-FOV CVGL task. We conducted a comprehensive evaluation with uniform training conditions. The results are summarised in Table 1. Although ConvNeXt has demonstrated promising results in previous work [2, 32], its performance on this challenging limited-FOV dataset is the weakest among all evaluated backbone architectures. This may be due to the limited capacity of ConvNeXt to capture sufficient contextual information from narrow FOV images. We then experiment with two large-scale backbones, ViT and DINOv2, observing that even with similar model structures, ViT [3] is consistently worse than DINOv2 [14]. Interestingly, with almost the same structure, DINOv2 consistently performs better than DINOv2 on Base scale ("B") and Large scale ("L"). This performance increase might result from 1) the pre-trained knowledge of DINOv2 on the LVD-142M dataset [14], and 2) the native input image size contains more fine-grained details (as illustrated by the different FLOPs). From this study, we selected DINOv2-L as the stage I backbone for VICI.

Drone Perspective Augmentation: The second experiment evaluates the effectiveness of the drone perspective augmentation, as summarised in Table 2. For each training sample, we design a probability P to replace the satellite image with a randomly sampled drone-view image from the same location. In this experiment, we set P to 0, 0.1, 0.3, and 0.5, respectively. As we can see, by setting P to 0.1 and 0.3, the model performance significantly improves - with $P = 0.3$ achieving the best results. However, with $P = 0.5$, performance drops significantly and is similar to the non-augmented case. Thus, we choose $P = 0.3$ for the probability of drone perspective augmentation during the training.

[1] https://ai.google.dev/gemini-api/docs/structured-output

Backbone	Params (M)	FLOPs (G)	Dims	R@1	R@5	R@10
ConvNeXt-T	28	4.5	768	1.36	4.34	7.95
ConvNeXt-B	89	15.4	1024	3.14	8.14	13.22
ViT-B	86	17.6	768	3.30	8.92	13.96
ViT-L	307	60.6	1024	9.62	23.42	32.73
DINOv2-B	86	152	768	17.37	36.14	46.96
DINOv2-L	304	507	1024	27.49	51.96	63.13

Table 1: Backbone capabilities evaluation.

P	R@1	R@5	R@10
0	24.47	48.16	60.99
0.1	26.98	51.34	61.92
0.3	27.49	51.96	63.13
0.5	24.89	52.03	62.66

Table 2: Drone augmentation with varying probability P.

VLM Re-ranking: The top 10 retrieved results for each query ground image are fed into stage II for VLM Re-ranking. To balance computational cost and efficiency, we choose two different variants of Google's Gemini 2.5 model - Gemini 2.5 Flash and Gemini 2.5 Flash Lite. We also fix the thinking budget at 1024 for both models to achieve the best efficiency. Results are summarised in Table 3. By comparing the results with and without re-ranking utilising Gemini 2.5 Flash, R@1 increases by 2.72% and R@5 increases by 1.08%, supporting the idea of leveraging VLMs to re-rank the coarse retrieving results. To further investigate the VLM re-ranking performance on small-scale models, we conducted the same experiment on the Gemini 2.5 Flash Lite. However, performance is worse than without re-ranking, illustrating that large-scale models with better reasoning capabilities play a critical role in this task.

VLM	R@1	R@5	R@10
Without Re-ranking	27.49	51.96	63.13
Gemini 2.5 Flash Lite	23.54	48.39	63.13
Gemini 2.5 Flash	30.21	53.04	63.13

Table 3: VLM re-ranking comparison.

Ablation and Summary of VICI: The ablation study of VICI and comparison with two previous state-of-the-art methods are stated in Table 4. As summarised, VICI substantially outperforms two baselines [26, 35] on the University-1652 dataset. Also, the proposed drone-view augmentation and VLM re-ranking demonstrate their effectiveness in improving the performance on this benchmark.

5 Conclusion

In this paper, we present VICI, a novel VLM-powered CVGL model that achieved outstanding performance on the UAVM 2025 challenge [25] without introducing any extra datasets. To boost the coarse localisation performance of limited FOV query images, we introduce the drone perspective augmentation strategy. Furthermore, we prove that the reasoning capability of existing foundational

Model	R@1	R@5	R@10
U1652 [35]	1.20	-	-
LPN w/o drone [26]	0.74	-	-
LPN w/ drone [26]	0.81	-	-
DINOv2-L	24.66	48.00	59.02
+ Drone Data	27.49	51.96	63.13
+ VLM Re-rank (Ours)	30.21	53.04	63.13

Table 4: Ablation study and baseline comparison.

VLMs significantly improves the localisation accuracy and provides fine-grained justifications for better interpretability, putting forward a new localisation paradigm for future research.

Acknowledgments

This work was supported by the National Science Foundation under Grants No. 2218063 and the EPSRC under grant agreement EP/S035761/1. Computations were supported in part by Advanced Micro Devices, Inc., under the AMD AI & HPC Cluster Program, and Vermont Advanced Computing Center (VACC).

References

[1] Barkin Dagda, Muhammad Awais, and Saber Fallah. 2025. GeoVLM: Improving Automated Vehicle Geolocalisation Using Vision-Language Matching. arXiv:2505.13669 [cs.CV] https://arxiv.org/abs/2505.13669

[2] Fabian Deuser, Konrad Habel, and Norbert Oswald. 2023. Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation. arXiv:2303.11851 [cs.CV]

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[4] Google Cloud. 2025. Gemini 2.5 Flash Model. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash. Accessed: 2025-06-15.

[5] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 2256–2264.

[6] Gaoshuang Huang, Yang Zhou, Luying Zhao, and Wenjian Gan. 2025. CV-Cities: Advancing Cross-View Geo-Localization in Global Cities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18 (2025), 1592–1606. doi:10.1109/JSTARS.2024.3502160

[7] Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. 2024. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Forty-first International Conference on Machine Learning*.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

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

[10] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11976–11986.

[11] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[12] Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. 2024. Congeo: Robust cross-view geo-localization across ground view variations. In *European Conference on Computer Vision*. Springer, 214–230.

[13] OpenAI. 2024. ChatGPT (July 5 Version). https://chat.openai.com/. Accessed: 2025-07-05.

[14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[16] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. 2020. Where Am I Looking At? Joint Location and Orientation Estimation by Cross-View Matching. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 4063–4071.

[17] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. 2020. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4064–4072.

[18] Tavis Shore, Oscar Mendez, and Simon Hadfield. 2025. PEnG: Pose-Enhanced Geo-Localisation. *IEEE Robotics and Automation Letters* 10, 4 (2025), 3835–3842. doi:10.1109/LRA.2025.3546513

[19] Tavis Shore, Oscar Mendez, and Simon Hadfield. 2025. SpaGBOL: Spatial-Graph-Based Orientated Localisation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 6858–6867. doi:10.1109/WACV61041.2025.00667

[20] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[21] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems* 37 (2024), 84839–84865.

[22] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[24] Tingyu Wang, Yujiao Shi, Fabian Deuser, Shaofei Huang, Guosheng Hu, Si Liu, Zhedong Zheng, and Roger Zimmermann. 2025. The 3rd Workshop on UAVs in Multimedia: Capturing the World from a New Perspective. In *Proceedings of the 33rd ACM International Conference on Multimedia Workshop*.

[25] Tingyu Wang, Yujiao Shi, Fabian Deuser, Shaofei Huang, Guosheng Hu, Si Liu, Zhedong Zheng, and Roger Zimmermann. 2025. The 3rd Workshop on UAVs in Multimedia: Capturing the World from a New Perspective. In *Proceedings of the 33rd ACM International Conference on Multimedia Workshop*.

[26] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. 2021. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 2 (2021), 867–879.

[27] Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. 2024. Image and object geo-localization. *International Journal of Computer Vision* 132, 4 (2024), 1350–1392.

[28] Scott Workman and Nathan Jacobs. 2015. On the location dependence of convolutional neural network features. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 70–78. doi:10.1109/CVPRW.2015.7301385

[29] Scott Workman, Richard Souvenir, and Nathan Jacobs. 2015. Wide-Area Image Geolocalization with Aerial Reference Imagery. In *IEEE International Conference on Computer Vision (ICCV)*. 1–9. doi:10.1109/ICCV.2015.451

[30] Hongji Yang, Xiufan Lu, and Ying J. Zhu. 2021. Cross-view Geo-localization with Layer-to-Layer Transformer. In *Neural Information Processing Systems*.

[31] Junyan Ye, Honglin Lin, Leyan Ou, Dairong Chen, Zihao Wang, Conghui He, and Weijia Li. 2024. Where am I? Cross-View Geo-localization with Natural Language Descriptions. *arXiv preprint arXiv:2412.17007* (2024).

[32] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. 2024. GeoDTR+: Toward Generic Cross-View Geolocalization via Geometric Disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 10419–10433. doi:10.1109/TPAMI.2024.3443652

[33] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. 2023. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 3480–3488.

[34] Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. 2023. Cross-view image sequence geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2914–2923.

[35] Zhedong Zheng, Yunchao Wei, and Yi Yang. 2020. University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization. *ACM Multimedia* (2020).

[36] Sijie Zhu, Mubarak Shah, and Chen Chen. 2022. TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 1152–1161.

[37] Sijie Zhu, Taojiannan Yang, and Chen Chen. 2021. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3640–3649.

[38] Yingying Zhu, Hongji Yang, Yuxin Lu, and Qiang Huang. 2023. Simple, Effective and General: A New Backbone for Cross-view Image Geo-localization. arXiv:2302.01572 [cs.CV]