Visual Recognition Spring 2025 Final Project – Image Matching Challenge 2023

Group 13 SIGGRAPH

June 2, 2025

1 Introduction

Problem statement. The Image Matching Challenge 2023 [5] aims to advance the field of 3D reconstruction from real-world image collections. In this competition, the goal is to estimate the six degrees of freedom (6DoF) camera poses for a collection of unordered images taken from diverse viewpoints and conditions. Given a batch of images without additional sensor data or structured metadata, participants are required to predict the relative rotation and translation of each image, enabling accurate 3D reconstruction of the scene via Structure from Motion (SfM) techniques.

The importance of this problem. Robust multi-view Structure-from-Motion (SfM) is a fundamental task in computer vision. It helps with things like search-and-rescue missions, self-driving navigation, preserving historical sites in digital form, and everyday services like street views in online maps. Google Maps, for example, relies on SfM to fuse crowdsourced photos into city-scale 3D models. As more and more people take photos with their smartphones, using this large amount of casual, shared photos could make it possible to create high-quality maps without needing specialized equipment. Therefore, solving this challenge will contribute to faster, more accessible, and more detailed reconstructions of the world around us.

The difficulties we address. The primary challenges lie in the diversity and unpredictability of real-world image data. Variations in viewpoint, lighting, weather, resolution, occlusion and noise make image matching highly challenging. Moreover, the absence of sequential or temporal context further increases the difficulty of finding robust correspondences. Another major challenge is scale and efficiency —real-world scenes can contain hundreds of images, and trying to match every pair of images quickly becomes computationally expensive and impractical. In this work, we aim to address these challenges by adopting learned feature extractors and robust pose estimation pipelines that generalize across diverse environments.

2 Related Works

Local Feature Extraction. Early Structure-from-Motion (SfM) methods used traditional feature descriptors like SIFT [15], which is still considered a reliable baseline today. However, the rise of learning-based descriptors has shifted the focus to methods such as SuperPoint [6], which uses synthetic training to simulate geometric variations. Other methods, like R2D2 [23] introduce robustness through local reliability scoring, while D2-Net [7] combines keypoint detection and description in a single network using features from convolution layers.

Feature Matching. Accurate matching is still a challenge in wide-baseline settings. SuperGlue [6] improves over MNN + RANSAC [8] by using self- and cross-attention over keypoints. It has become the dominant matcher in academic and competition settings. LoFTR [30] extends this by bypassing detection entirely, offering dense match prediction even in textureless or repetitive regions. LightGlue [13] further reduces computation time while keeping the same high quality, making it ideal for batch reconstruction of more images.

Pose Estimation & SfM. COLMAP [26] remains the most widely used SfM system, combining exhaustive or vocabulary-based matching with incremental reconstruction and non-linear bundle adjustment. Alternatives such as OpenMVG [21] and Theia [31] support global SfM, solving for camera rotations and translations via rotation averaging and pose graph optimization before 3D point triangulation.

Feed-forward reconstruction. Recent advances have introduced a paradigm shift from traditional optimization-based SfM to direct regression approaches. DUSt3R [37] pioneered pointmap regression without camera calibration, unifying monocular and binocular cases through Transformers. MASt3R [10] enhanced this with dense local features and fast reciprocal matching for extreme viewpoint changes. VGGSfM [34] created the first fully end-to-end differentiable SfM pipeline, integrating deep 2D tracking with differentiable bundle adjustment. VGGT [36] represents the latest unified solution, directly inferring all 3D scene attributes from single or multiple views in seconds without post-processing optimization.

3 Method

3.1 Baseline: `KeyNetAffNetHardNet` + COLMAP

Image Pair Selection Global image descriptors are extracted using a pre-trained `tf_efficientnet_b7` model. Image pairs are shortlisted based on descriptor cosine similarity using a threshold (e.g., 0.45) or by selecting a minimum number of top matches (e.g., 35 pairs), with exhaustive pairing for small image sets (e.g., < 20 images).

Local Feature Extraction and Matching For shortlisted pairs, `KeyNetAffNetHardNet` (Kornia) extracts approximately 2048 upright local features per image, after resizing the smaller image edge to 600 pixels. These keypoints (LAFs) and HardNet descriptors are saved to HDF5 files. Symmetric Mutual Nearest Neighbor (SMNN) matching is then performed, with results also stored in HDF5.

COLMAP Database Import A custom SQLite interface populates a COLMAP database with camera information (from EXIF or prior, using a 'simple-radial' model), image details, keypoints, and putative matches from the HDF5 files.

3D Reconstruction with PyCOLMAP Geometric verification via `pycolmap.match_exhaustive()` (RANSAC) precedes 3D scene reconstruction and pose estimation using `pycolmap.incremental_mapping()`. Mapper options permit model generation with 3 registered images. The reconstruction with the most registered images is selected, and its camera rotation matrices (R) and translation vectors (t) are formatted for submission.

3.2 Method 1: Ensemble Matching

This method builds upon the 7th place solution from the Image Matching Challenge 2023 [5]. The pipeline integrates efficient image retrieval, diverse local feature extractors, the novel deep matcher LightGlue [13], and robust SfM modules. By ensembling complementary techniques across different stages, the method aims to improve both robustness and accuracy in large-scale image matching and 3D reconstruction tasks.

Image Retrieval. To reduce computational cost due to Kaggle restrictions, NetVLAD [2] was introduced as a global descriptor for image retrieval. It enables compact yet discriminative representations of images, making it suitable for large-scale scenes. For each query images, the top-K most similar images are selected to form candidate pairs.

Feature Extraction. We utilized an ensemble of keypoint detectors and local feature descriptors to leverage their individual strengths and promote feature diversity. It increases the chance of detecting repeatable and discriminative keypoints under varying conditions such as illumination, scale, and viewpoint changes:

- ALIKED [39]: A lightweight CNN-based feature extractor with deformable convolutions.
- DISK [33]: A reinforcement-learning based local feature extractor with up to 5000 keypoints per image.
- SIFT [15]: A classical method with strong rotation invariance, particularly beneficial in heritage scenes with geometric transformations.

Feature Matching. We adopted LightGlue [13], a fast and accurate transformer-based matcher trained on the MegaDepth dataset [12]. It was applied to features extracted from ALIKED [39], DISK [33]. Instead of using the default configuration, we adjust the filtering threshold to achieve better matching performance. For SIFT [15] features, we employed traditional nearest-neighbor ratio matching.

SfM & Localizing Unregistered Images. The resulting feature matches are passed to PixSfM [14], a SfM framework that incorporates learned priors and photometric refinement into the standard SfM pipeline. Finally, for images that were not registered in the initial reconstruction, we used the HLoc toolbox [24, 25] to estimate their poses. These steps ensured more complete reconstructions and improved scoring coverage.

3.3 Method 2: Multi-Expert Dense & Sparse Fusion

We proposed a hybrid sparse-dense SfM pipeline that merges classical keypoints with MASt3R grid matches [11], filters them in geometry space, and reconstructs the scene in one COLMAP pass [27]. All modules are designed to finish within the Kaggle nine-hour GPU limit.

Image pair selection. Every input image is resized so that its longer edge equals 224 pixels, then a DINOv2 ViT-B/14 encoder produces a global embedding [22]. Cosine distances between all embeddings form a similarity matrix. Pairs whose distance is at most 0.3 are accepted, and the nearest neighbours are appended until each view appears in at least 50 pairs. This strategy guarantees full graph connectivity while avoiding the quadratic cost of exhaustive pairing.

Image rotation preprocessing. Before feature extraction every image passes through `process_image_rotation`, a lightweight orientation-classification step based on a SWSL-ResNeXt-50 model [17] trained for four discrete angles $\{0°, 90°, 180°, 270°\}$. The predicted angle determines an in-place rotation that restores upright geometry, so all downstream detectors and dense matchers operate on consistently oriented inputs. The chosen angle is cached and later reapplied inversely to the recovered camera poses, ensuring that the final reconstruction remains in the original coordinate frame.

Sparse feature pipeline. After the rotation–normalisation step every upright image is resized so that its longer edge equals 1600 pixels. We then run four complementary keypoint detectors, each followed by OriNet [4] for orientation, AffNet [19] for affine shape, and HardNet-8 [18] for description. Using multiple detectors increases the probability of finding repeatable and discriminative keypoints under diverse illumination, scale, and viewpoint changes.

- **KeyNet** [4] learns scale-space extrema directly from data and provides well-distributed interest points.
- **Good–Features–to–Track** [28] detects strong corners through the minimum eigenvalue criterion and performs reliably on low-texture regions.
- **Difference-of-Gaussians** [16] approximates Laplacian-of-Gaussian extrema, offering robustness to scale variations and blur.
- **Harris corners** [9] supplies inexpensive yet rotation-aware points that bolster detector diversity in highly textured areas.

HardNet-8 descriptors are sampled on 32-pixel patches centred at each detected keypoint. For every selected image pair AdaLAM [32] establishes tentative correspondences and removes local affine outliers in a single GPU pass.

Dense feature pipeline. MASt3R complements the sparse tracks with grid-based correspondences. Each image is resized so that its longer edge equals 512 pixels and is sampled on a 16-pixel grid. Symmetric inference computes descriptors for both directions of every selected pair. Points whose descriptor confidence is at least 1.5 survive, and the grid indices convert back to original coordinates by the stored scale factors.

Match consolidation and geometric filtering. Sparse and dense matches are concatenated, then duplicate correspondences are removed by hashing the two-dimensional coordinates. MAGSAC-USAC [3] estimates the fundamental matrix for every pair with a reprojection threshold of 5 pixels, a confidence of 99.9 %, and 50 000 iterations. Matches inconsistent with the recovered model are discarded, and pairs whose inlier count is lower than 1 % of the scene maximum are ignored in later stages.

Structure-from-motion. All surviving keypoints, descriptors, matches, and fundamentals are written to an SQLite database that follows the COLMAP schema. An incremental mapper from the pycolmap library reconstructs the scene. Key parameters are a minimum triangulation angle of two degrees, a maximum reprojection error of three pixels, a minimum track length of three observations, and a local bundle-adjustment angle floor of eight degrees. Absolute pose registration requires at least thirty inliers with an angular error under twelve degrees. Parallel bundle adjustment activates automatically once more than forty views are registered.

3.4 Method 3: VGGT [36] with BA

Camera Pose Estimation with VGGT. This method leverages the state-of-the-art feed-forward reconstruction model, VGGT [36], to estimate camera poses, using the demo script provided by the paper authors as a starting point[1].

Initial Scene Reconstruction. For each image, VGGT predicts camera parameters (extrinsic and intrinsic) and a depth map. A pointmap is then generated by projecting image pixels into world space using these camera parameters and the depth map. The comprehensive 3D point cloud is subsequently formed by fusing all individual pointmaps.

Track Generation with VGGSfM. For efficiency, track prediction incorporates elements from VGGSfM [35] instead of VGGT's native track predictor. This process begins by sampling N query frames via farthest point sampling (FPS), where distances are based on the cosine similarity of DINOv2 features to ensure distinctive image selection. For each query frame, the previously generated 3D point cloud is projected onto it. Only 2D image points that correspond to these 3D points are considered keypoints. Given this set of keypoints, VGGSfM's track predictor then generates tracks that match these keypoints.

Final Refinement. The camera parameters, 3D point cloud, generated tracks, and keypoints collectively form an initial 3D reconstruction. This reconstruction is then further refined using COLMAP's bundle adjustment algorithms to optimize the overall scene geometry and camera poses.

Performance and Limitations. While this method utilizes the powerful VGGT model and integrates VGGSfM's efficient track predictor to reduce inference time, memory limitations become a significant concern, as indicated in Table 1. Processing more than 100 input frames typically leads to out-of-memory errors on platforms like the Kaggle notebook's 16 GB GPU. This constraint is exacerbated by the P100 GPU's lack of `bfloat16` support, necessitating the use of higher-precision data types which increases memory demands. Our experiments show that even a small dataset of 15 images consumes approximately 15.6 GB of GPU memory. Therefore, this method was evaluated on an A6000 GPU equipped with 48 GB of memory to overcome these limitations.

[1] `https://github.com/facebookresearch/vggt/blob/main/demo_colmap.py`

Table 1: **Runtime and peak GPU memory usage across different numbers of input frames.** Runtime is measured in seconds, and GPU memory usage is reported in gigabytes. This table is adopted from VGGT [36] paper.

Input Frames	**1**	**2**	**4**	**8**	**10**	**20**	**50**	**100**	**200**
Time (s)	0.04	0.05	0.07	0.11	0.14	0.31	1.04	3.12	8.75
Mem. (GB)	1.88	2.07	2.45	3.23	3.63	5.58	11.41	21.15	40.63

Table 2: **Train score.**

Method	bike	chairs
Multi-Expert Dense & Sparse Fusion	0.918	**0.979**
VGGT [36] with BA	**0.922**	0.854
Ensemble Matching	**0.922**	0.801

Scalable Reconstruction. To address the aforementioned memory limitation, we propose a divide-and-conquer approach inspired by scalable SfM literature [1, 38, 20, 29]. Our method splits input images into B overlapping batches, reconstructs each batch independently using the same VGGT pipeline, and merges the B partial reconstructions into a complete model using `colmap model_merger`. This approach aims to leverage VGGT's capabilities while circumventing memory constraints. However, implementing this pipeline presents significant technical challenges. The `model_merger` requires identical images across different reconstructions to share consistent feature points (`point2D`) and correspondence tracks. This necessitates maintaining a global database to track 3D points, keypoints, and their associations across all partial reconstructions—a non-trivial engineering challenge involving careful management of feature consistency and geometric transformations between overlapping batches. While we were unable to complete this implementation within the project timeline, we believe this represents a promising direction for future work that could effectively combine the advantages of modern feed-forward reconstruction with established scalable SfM principles.

4 Experiments

We show the comparisons of of different methods on the Image Matching Challenge 2023 in Tab. 3. We report both public and private leaderboard scores. *Ensemble Matching (modified)*, which incorporates a tuned filtering threshold, achieves the highest scores on both leaderboards, indicating strong generalization and robust performance across diverse scenes.

Additional Experiments. Due to the OOM issue of VGGT with BA, we test it on the public scene dataset compared to the Ensemble Matching. The results are shown on the Tab. 2 and it demonstrates the possible future solution of our VGGT with BA.

5 Final ranking

Our public and private scores, as shown in Fig. 1, were 0.483 and 0.548, respectively. Based on the private leaderboard results, Fig. 2, our method achieved a top-3 ranking, surpassing the strong baseline.

6 Conclusion

Our experiments demonstrate that this multi-expert ensemble approach yields competitive performance, achieving a top-3 rank on the private leaderboard with a score of 0.548, which surpassing strong baselines and illustrating the effectiveness of combining both classical and learning-based components. This challenge highlights the importance of both accurate matching and completeness in 3D reconstruction pipelines. Our conclusions suggest that future improvements can be made by further optimizing feature diversity, dynamic pair selection, and adaptive matching strategies.

7 GitHub Link

`https://github.com/jayin92/NYCU-VRDL-final-project`

References

[1] Sameer Agarwal et al. "Building Rome in a day". In: *Communications of the ACM*. Vol. 54. 10. 2011, pp. 105–112.

[2] Relja Arandjelovic et al. "NetVLAD: CNN architecture for weakly supervised place recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016, pp. 5297–5307.

[3] Daniel Barath et al. "MAGSAC++: A Fast, Reliable and Accurate Robust Estimator". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021, pp. 1304–1312.

[4] Axel Barroso-Laguna et al. "Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2019, pp. 118–127.

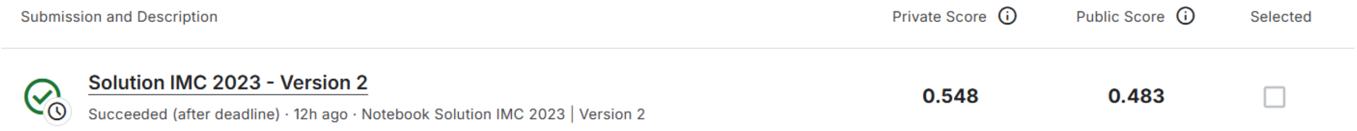
[5] Ashley Chow et al. *Image Matching Challenge 2023.* `https://kaggle.com/competitions/image-matching-challenge-2023`. Kaggle. 2023.

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. "SuperPoint: Self-supervised interest point detection and description". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops.* 2018, pp. 224–236.

[7] Mihai Dusmanu et al. "D2-Net: A trainable CNN for joint detection and description of local features". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2019, pp. 8092–8101.

Table 3: **Experimental results.**

Method	Public Score	Private Score
`KeyNetAffNetHardNet` + COLMAP	0.160	0.174
Multi-Expert Dense & Sparse Fusion	0.465	0.515
VGGT [36] with BA	-	-
Ensemble Matching (default)	0.465	0.494
Ensemble Matching (modified)	**0.483**	**0.548**



Submission and Description	Private Score	Public Score	Selected
Solution IMC 2023 - Version 2 Succeeded (after deadline) · 12h ago · Notebook Solution IMC 2023 | Version 2	**0.548**	**0.483**	□

Figure 1: Final scores.

[8] Martin A. Fischler and Robert C. Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Commun. ACM* 24.6 (June 1981), pp. 381–395. ISSN: 0001-0782. DOI: `10.1145/358669.358692`. URL: `https://doi.org/10.1145/358669.358692`.

[9] Chris Harris and Mike Stephens. "A Combined Corner and Edge Detector". In: *Proceedings of the Alvey Vision Conference*. 1988, pp. 147–151.

[10] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. "Grounding Image Matching in 3D with MASt3R". In: *Computer Vision – ECCV 2024*. Springer, 2025, pp. 65–84.

[11] Vincent Leroy et al. "MASt3R: Matching by Attending to Scene Layout and Repetitions". In: *arXiv preprint arXiv:2401.09217* (2024).

[12] Zhengqi Li and Noah Snavely. "MegaDepth: Learning Single-View Depth Prediction from Internet Photos". In: *Computer Vision and Pattern Recognition (CVPR)*. 2018.

[13] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. "LightGlue: Local feature matching at light speed". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 17627–17638.

[14] Philipp Lindenberger et al. "Pixel-perfect structure-from-motion with featuremetric refinement". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 5987–5997.

[15] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[16] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110.

[17] Dhruv Mahajan et al. "Exploring the Limits of Weakly Supervised Pretraining". In: *European Conference on Computer Vision*. 2018, pp. 185–201.

[18] Anastasiia Mishchuk et al. "Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss". In: *Advances in Neural Information Processing Systems*. 2017, pp. 4829–4838.

[19] Dmytro Mishkin, Filip Radenovic, and Ondrej Chum. "Repeatability Is Not Enough: Learning Affine Regions via Discriminability". In: *European Conference on Computer Vision*. 2018, pp. 284–300.

[20] Pierre Moulon, Pascal Monasse, and Renaud Marlet. "Global fusion of relative motions for robust, accurate and scalable structure from motion". In: (2013), pp. 3248–3255.

[21] Pierre Moulon et al. "OpenMVG: Open multiple view geometry". In: *International Workshop on Reproducible Research in Pattern Recognition*. Springer. 2016, pp. 60–74.

[22] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2023. arXiv: `2304.07193 [cs.CV]`.

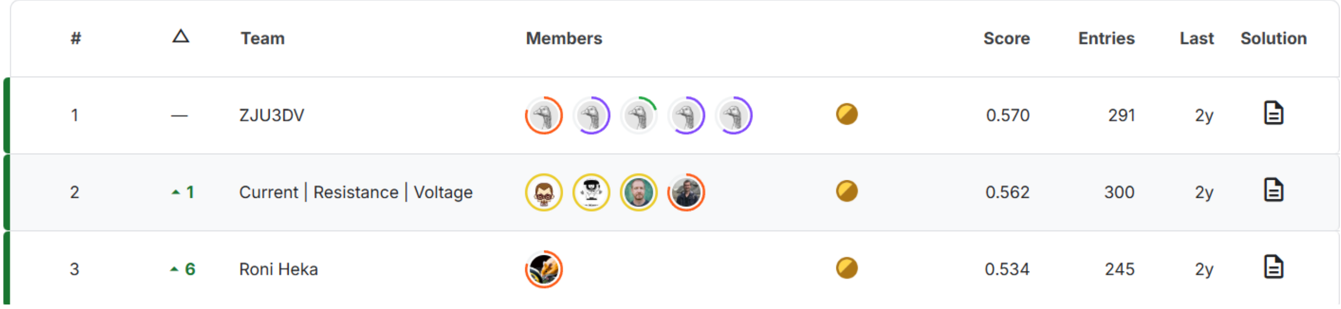
[23] Jérôme Revaud et al. "R2D2: Reliable and repeatable detector and descriptor". In: *Advances in neural information processing systems*. Vol. 32. 2019.

[24] Paul-Edouard Sarlin et al. "From Coarse to Fine: Robust Hierarchical Localization at Large Scale". In: *CVPR*. 2019.

[25] Paul-Edouard Sarlin et al. "SuperGlue: Learning feature matching with graph neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 4938–4947.

[26] Johannes L Schönberger and Jan-Michael Frahm. "Structure-from-motion revisited". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4104–4113.

[27] Johannes L. Schönberger and Jan-Michael Frahm. "Structure-from-Motion Revisited". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4104–4113.



#	△	Team	Members		Score	Entries	Last	Solution
1	—	ZJU3DV		●	0.570	291	2y	🗎
2	▲ 1	Current | Resistance | Voltage		●	0.562	300	2y	🗎
3	▲ 6	Roni Heka		●	0.534	245	2y	🗎

Figure 2: Private leaderboard score.

[28] Jianbo Shi and Carlo Tomasi. “Good Features to Track”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1994, pp. 593–600.

[29] Noah Snavely, Steven M Seitz, and Richard Szeliski. “Skeletal graphs for efficient structure from motion”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.

[30] Jiaming Sun et al. “LoFTR: Detector-free local feature matching with transformers”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 8922–8931.

[31] Chris Sweeney. *Theia Multiview Geometry Library: Tutorial & Reference*. `http://theia-sfm.org`.

[32] Mykhailo Tsybulya et al. “AdaLAM: Revisiting Handcrafted Outlier Rejection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 9263–9272.

[33] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. “DISK: Learning local features with policy gradient”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 14254–14265.

[34] Jianyuan Wang et al. “VGGSfM: Visual Geometry Grounded Deep Structure From Motion”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 21686–21697.

[35] Jianyuan Wang et al. “VGGSfM: Visual Geometry Grounded Deep Structure From Motion”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 21686–21697.

[36] Jianyuan Wang et al. “VGGT: Visual Geometry Grounded Transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2025.

[37] Shuzhe Wang et al. “DUSt3R: Geometric 3D Vision Made Easy”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 20697–20709.

[38] Kyle Wilson and Noah Snavely. “Robust global translations with 1dsfm”. In: *European Conference on Computer Vision*. Springer, 2014, pp. 61–75.

[39] Xiaoming Zhao et al. “Aliked: A lighter keypoint and descriptor extraction network via deformable transformation”. In: *IEEE Transactions on Instrumentation and Measurement* 72 (2023), pp. 1–16.