Deep Dive into Feature Matching under Large Viewpoint Changes

Yin-Kuan Lee[1], Chun-An Chen[1], and Yi-Cheng Wang[1]

National Taiwan University, Taipei 10617, Taiwan
r14946012@g.ntu.edu.tw, cyi871208@gmail.com, wang.amiannn@gmail.com

Abstract. In this project, we investigate how advanced feature matching models, specifically MAST3R and VGGT, behave under large viewpoint changes and aim to better understand the practical limitations of current methods. To support this study, we constructed a challenging dataset consisting of 1,778 real-world image pairs from Mapfree, and additionally designed 20 special-case pairs via controlled image editing to stress-test specific conditions. Extensive experiments show a clear degradation in performance as viewpoint differences increase, with AUC@20 dropping from 60 to 35, indicating that large-angle changes remain a fundamental challenge. Beyond quantitative trends, our qualitative analysis offers insight into failure behavior: models struggle particularly with unseen, highly reflective, and rounded objects, where geometric cues weaken or become distorted. We also observe frequent reliance on contextual structures such as shadows, trees, and railings to maintain correspondences, suggesting that background context plays a more important role than commonly assumed. Overall, these findings reveal critical limitations of current state-of-the-art feature matching models and provide valuable guidance for designing more robust and viewpoint-invariant matching systems in the future.

Keywords: Feature Matching · Large Viewpoint Changes · 3D Reconstruction · MAST3R · VGGT.

1 Introduction

3D reconstruction from sparse views is highly desirable for its efficiency compared to dense view reconstruction. However, sparse views often imply large viewpoint changes, which pose significant challenges for feature matching algorithms. Large changes in viewpoint lead to severe perspective distortions, scale variations, and occlusions, making it difficult for traditional local feature matchers to find reliable correspondences. While recent data-driven models like MAST3R [1] and VGGT [2] incorporate strong geometric priors and transformer-based architectures to handle such changes, their performance limits under extreme conditions—such as viewpoint changes exceeding 90 degrees—are not fully understood.

Our goal is to:

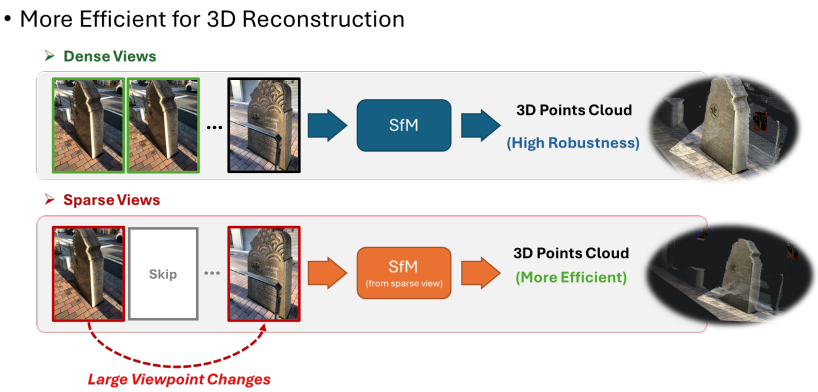

Fig. 1. Motivation: Comparison between Dense Views (top) and Sparse Views (bottom). Sparse views are more efficient for 3D reconstruction but introduce large viewpoint changes, making feature matching more challenging.

1. Explore how different models handle large viewpoint changes, specifically focusing on the degradation of performance as the angle increases.
2. Understand the limitations of current state-of-the-art methods (MAST3R [1], VGGT [2]) through both quantitative metrics and qualitative failure analysis.

Our main contributions are:

- We built a large view changes dataset containing 1,778 pairs filtered from the Mapfree dataset [3], specifically targeting wide-baseline scenarios.
- We created 20 pairs of edited images to test special cases (changing objects, textures, backgrounds) to perform a "real-world ablation study."
- We conducted extensive experiments showing performance degradation with increasing view angles, with AUC@20 dropping significantly.
- We identified specific failure modes, such as smooth/reflective objects, and found that models often rely on environmental context like shadows and vegetation rather than the object itself.

2 Dataset Collection

2.1 Large View Changes Dataset

We utilized the Mapfree dataset [3] as our source, which contains diverse outdoor scenes. To focus on our problem setting, we filtered the dataset based on the degree of viewpoint change and the physical distance between frames. We specifically targeted pairs with viewpoint changes between 45° and 180°. After automated filtering, we manually selected pairs to ensure visual overlap and remove completely occluded cases, resulting in a final dataset of 1,778 pairs.

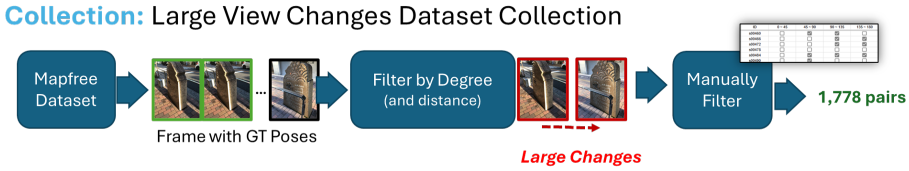

Fig. 2. Large View Changes Dataset Collection: We filtered the Mapfree dataset based on viewpoint change degree and distance, followed by manual selection to ensure high-quality pairs.

2.2 Special Cases Dataset

To analyze specific failure modes and perform a controlled study, we created a smaller dataset of 20 pairs. We used the "Nano Banana" image editing tool to modify specific attributes of the scene, such as changing the object's texture, removing background elements, or altering lighting conditions. To ensure the edited images remained geometrically consistent for valid evaluation, we processed them through "Veo 3" (I2V) for consistency checks. This allowed us to isolate factors affecting model performance, such as texture vs. geometry.

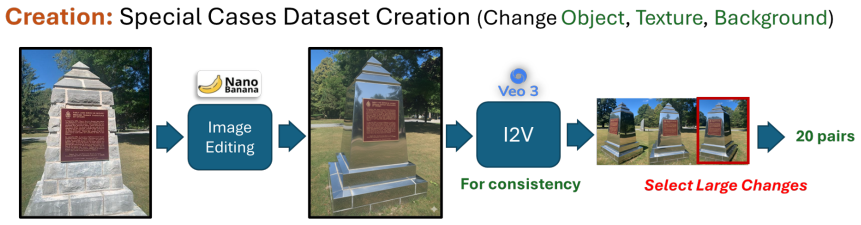

Fig. 3. Special Cases Dataset Creation: We employed image editing tools (I2V) to modify specific attributes (e.g., object texture, background) while maintaining geometric consistency, creating 20 pairs for targeted testing.

3 Experiments

We evaluated MAST3R [1] and VGGT [2] on our collected dataset. We measured the pose estimation accuracy derived from feature matching results. The primary metric is the Area Under the Curve (AUC) of the pose error at thresholds of 5°, 10°, and 20°.

3.1 Relative Pose Evaluation Metrics

We evaluate relative pose estimation accuracy using rotation and translation angular errors, following standard practice in relative camera pose evaluation.

Rotation Error. Given the ground-truth relative rotation $\mathbf{R}_{ij}^{\mathrm{gt}} \in SO(3)$ and the estimated rotation $\hat{\mathbf{R}}_{ij}$, the rotation error is defined as

$$\Delta\theta_{\mathrm{rot}} = \arccos\left(\frac{\mathrm{tr}\left(\mathbf{R}_{ij}^{\mathrm{gt}^{\top}}\hat{\mathbf{R}}_{ij}\right) - 1}{2}\right), \tag{1}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace. The resulting angle is reported in degrees.

Translation Error. Let $\mathbf{t}_{ij}^{\mathrm{gt}} \in \mathbb{R}^3$ and $\hat{\mathbf{t}}_{ij}$ denote the ground-truth and estimated relative translation vectors, respectively. Since the scale of translation is ambiguous, we measure the angular error between translation directions:

$$\Delta\theta_{\mathrm{trans}} = \arccos\left(\frac{\hat{\mathbf{t}}_{ij}^{\top}\mathbf{t}_{ij}^{\mathrm{gt}}}{\|\hat{\mathbf{t}}_{ij}\|\,\|\mathbf{t}_{ij}^{\mathrm{gt}}\|}\right). \tag{2}$$

If either translation vector has near-zero magnitude, the translation error is set to zero. All translation errors are reported in degrees.

Combined Pose Error. For each image pair, we define a single pose error by taking the maximum of rotation and translation errors:

$$\varepsilon_i = \max\left(\Delta\theta_{\mathrm{rot}}, \Delta\theta_{\mathrm{trans}}\right). \tag{3}$$

Recall. Given a threshold τ, the recall is defined as the fraction of image pairs whose combined error is below τ:

$$\mathrm{Recall}(\tau) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\varepsilon_i < \tau\right], \tag{4}$$

where N is the total number of evaluated pairs and $\mathbb{I}[\cdot]$ is the indicator function.

Area Under the Recall Curve (AUC). To summarize performance across error thresholds, we compute the Area Under the Recall Curve (AUC) up to a maximum threshold $\tau_{\max}$:

$$\mathrm{AUC@}\tau_{\max} = \frac{1}{\tau_{\max}}\int_{0}^{\tau_{\max}}\mathrm{Recall}(\tau)\,d\tau. \tag{5}$$

In practice, the integral is approximated using the trapezoidal rule with dense sampling at 0.1° intervals. We report AUC@5°, AUC@10°, and AUC@20°.

3.2 Quantitative Analysis

Table 1 shows the AUC scores for different ranges of viewpoint changes. The dataset was binned into three categories: 45°–90°, 90°–135°, and 135°–180°.

Table 1. Performance comparison (AUC) under different viewpoint change degrees.

Degree	AUC@5°		AUC@10°		AUC@20°	
	MAST3R	VGGT	MAST3R	VGGT	MAST3R	VGGT
45°–90°	67.83	10.37	78.94	34.78	85.43	61.67
90°–135°	47.24	1.88	60.91	15.46	70.18	47.85
135°–180°	36.98	0.81	47.75	6.55	55.80	31.90

As observed, the error increases (AUC decreases) significantly as the viewpoint angle increases. For instance, MAST3R's AUC@20 drops from 85.43 in the 45°–90° bin to 55.80 in the 135°–180° bin. VGGT shows a similar trend but with generally lower performance, suggesting that MAST3R's architecture is more robust to these extreme changes. However, both models struggle significantly in the highest angle bin.

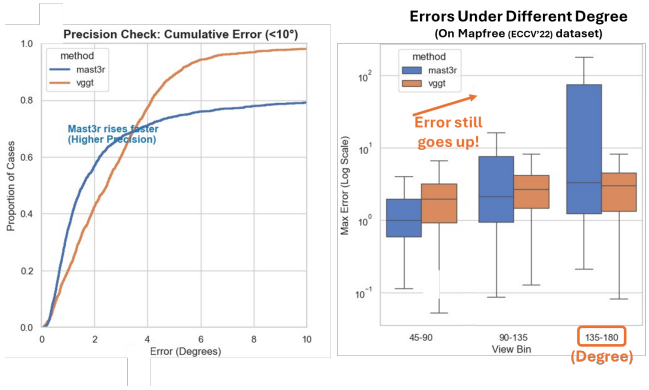

Fig. 4. Quantitative performance under different viewpoint changes. (Left) Cumulative pose error distribution shows a clear degradation trend as the viewpoint gap increases. (Right) Error statistics across different angle bins indicate significantly higher error in the 135°–180° range, highlighting the persistent difficulty of extreme viewpoint changes.

3.3 Qualitative Analysis and Case Studies

We analyzed the top 24 successful and failure cases to understand model behavior.

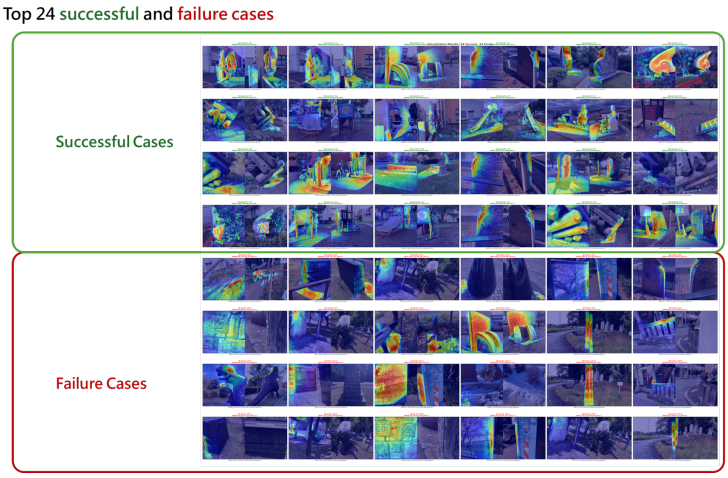

Fig. 5. Qualitative analysis of success and failure cases. Top row: Successful matching despite large viewpoint change. Bottom row: Failure case due to reflective surfaces.

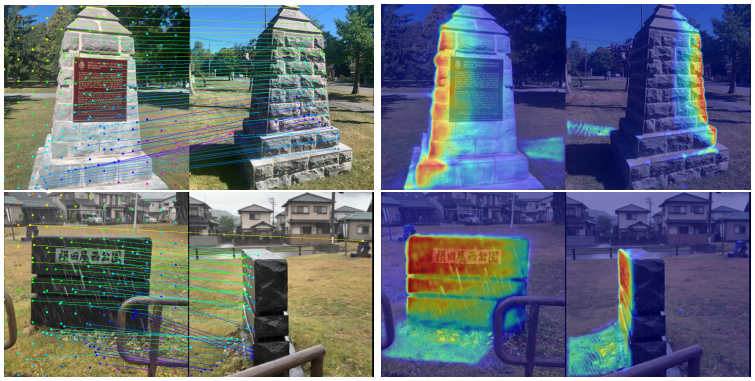

Fig. 6. Detailed Successful Cases: Examples where models successfully matched features despite significant viewpoint changes, often relying on strong geometric priors or distinct features.

Fig. 7. Failure Cases: Challenges with lighting variations, specific object properties, unknown objects.

Success Factors

- **Stable Structural Geometry:** Successful cases commonly involve scenes with strong, rigid geometric structures (e.g., sharp edges, planar stone surfaces, and well-defined corners), enabling the model to preserve consistent correspondences even under large viewpoint changes.
- **Texture and Surface Detail:** Distinct textures and engraved patterns (such as stone inscriptions) provide reliable local features, which significantly improve matching stability.
- **Lighting and Shadow Cues:** Lighting variations and shadow boundaries introduce additional structural cues, helping the model maintain correspondences when purely geometric information becomes insufficient.
- **Environmental Support Cues:** Surrounding context, including railings, ground planes, and background buildings, often serves as supplementary anchors, further stabilizing pose estimation when the target object alone is challenging.

Failure Modes

- **Challenging Object Properties:** Models frequently fail on smooth, rounded, and highly reflective surfaces, where stable local features are limited and appearance changes drastically with viewpoint.
- **Unfamiliar or Ambiguous Shapes:** Objects with unusual geometry (e.g., the fish-shaped slide) introduce ambiguity, making correspondence estimation unreliable.
- **Environmental and Illumination Conditions:** Low-light scenes, nighttime environments, and cluttered surroundings further degrade matching quality, indicating sensitivity to illumination and scene complexity.

Case Studies We performed ablation-like studies by editing images to isolate specific factors. This allows us to understand what features the models are actually relying on.

1. **Lighting and Shadows:** We altered the lighting configuration to change the position of cast shadows in the scene. As shown in Fig. 8, the model shifts its correspondences toward the new shadow boundaries, even though the underlying geometry remains unchanged. This behavior indicates that the model relies on transient shadow cues rather than purely on object geometry, suggesting that its matching process is not truly illumination-invariant.
2. **Background (Context):** We removed surrounding background elements such as trees and railings (Fig. 9). After this removal, the matching either became unstable or failed, indicating that when the primary object is difficult to match (e.g., surfaces lacking distinctive local features), the model relies heavily on contextual structures in the environment to establish correspondences.

3. **Texture:** We modified a smooth object by replacing its surface with a stone-like texture (Fig. 10). After adding texture, the correspondence became noticeably more stable, suggesting that textured surfaces provide rich local features that are difficult to obtain from purely smooth geometry.
4. **Unknown Object:** We introduced an object with an unusual shape (a fish-shaped slide) that is likely outside the training distribution (Fig. 11). The model failed to match this object effectively, highlighting the limitation of data-driven methods when encountering out-of-distribution geometries.

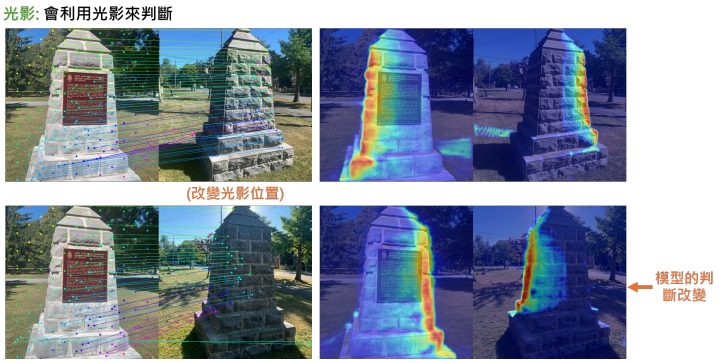

Fig. 8. Case Study: Lighting. Changing the light source position affects the model's matching, indicating reliance on shadows. Right: Keypoint matching, Left: Heatmap.

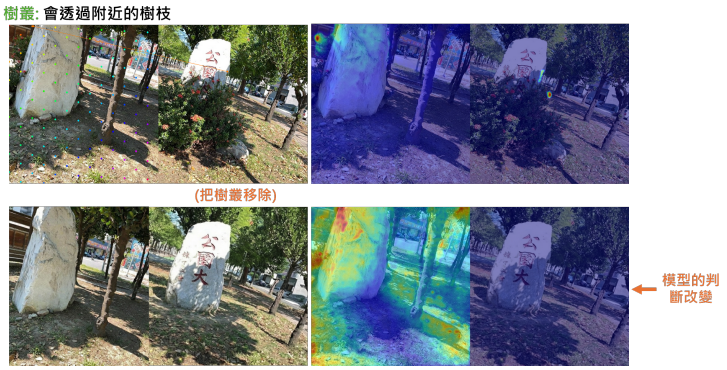

Fig. 9. Case Study: Background. Removing background elements like trees can lead to matching failure, showing the importance of contextual cues. Right: Keypoint matching, Left: Heatmap.

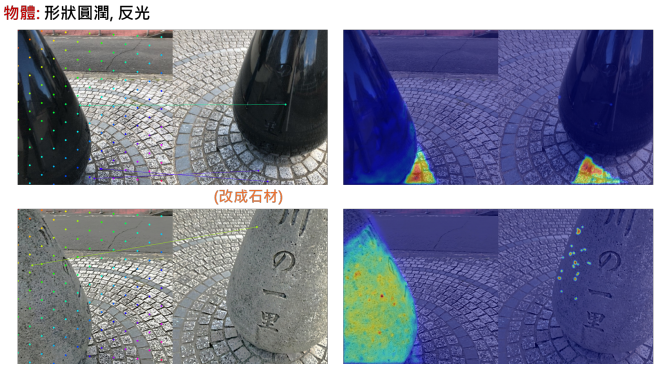

Fig. 10. Case Study: Texture. Changing the material of a smooth object to stone can improve matching performance, demonstrating the impact of texture. Right: Keypoint matching, Left: Heatmap.

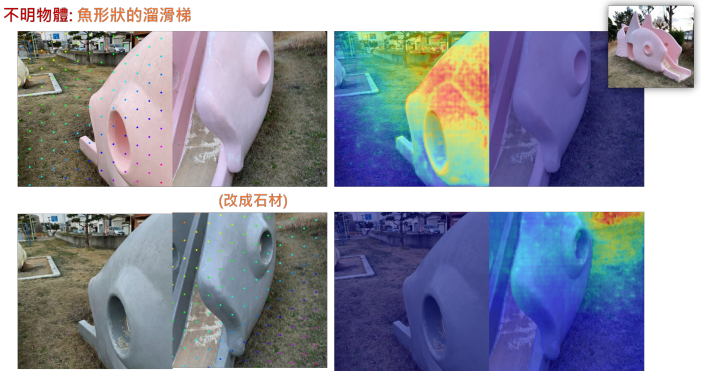

Fig. 11. Case Study: Unknown Object. Unusual objects, such as a fish-shaped slide, present significant challenges for current models. Right: Keypoint matching, Left: Heatmap.

4 Conclusion

In this work, we presented a comprehensive analysis of feature matching under large viewpoint changes, supported by a challenging dataset of 1,778 real-world image pairs and 20 carefully designed special cases. Our quantitative evaluation shows that current state-of-the-art models still experience substantial pose estimation degradation as viewpoint differences increase, with AUC dropping from 60 to 35. A head-to-head comparison further reveals that catastrophic failures become more frequent at the largest viewpoint ranges (135–180°), indicating that extreme viewpoint changes remain a largely unsolved problem.

Beyond global statistics, our qualitative analysis helps explain why these failures occur. Successful cases typically involve rigid structures with stable geometric edges and rich texture, enabling consistent correspondences. In contrast, failures are often linked to reflective, smooth or rounded, or appearance-changing objects, where geometric cues become ambiguous and photometric evidence unstable. Heatmap visualizations also show that models frequently rely on contextual background elements—such as shadows, trees, fences, and ground planes—rather than consistently focusing on the target object itself, suggesting a strong dependence on environmental cues.

Overall, these findings highlight two key challenges for future research: (1) improving robustness to object-intrinsic properties such as reflection, smooth curvature, and unseen surfaces; and (2) reducing reliance on transient contextual cues that may not be available or consistent across viewpoints. Addressing these issues will be essential for achieving reliable feature matching and pose estimation in truly unconstrained outdoor environments.

Lessons Learned

Through this project, we learned that state-of-the-art feature matching models are still heavily influenced by contextual cues and photometric factors, rather than relying purely on geometric consistency. We also realized the importance of analyzing model behavior beyond raw accuracy scores; qualitative case studies provided critical insight into why models fail, what factors truly drive matching stability, and where future improvements are most needed.

Work Division

- **Overall Collaboration:** The project was completed through close teamwork. All members jointly discussed methodology, interpreted results, refined the experimental design, and continuously improved the system and analyses throughout the entire project.
- **Yin-Kuan Li:**
  - Took charge of running the MAST3R experiments
  - Performed detailed quantitative evaluation and comparative analysis
  - Produced key performance figures, precision curves, and error distribution visualizations that supported major conclusions
  - Contributed to the written proposal report and actively participated in interpreting experimental findings
  - Assisted in refining and polishing the final written report
- **Yi-Cheng Wang:**
  - Led VGGT experiments and Mapfree pair selection
  - Refined and integrated the final AUC computation and organized the quantitative tables
  - Finalized and optimized the Special Case Dataset
  - Served as the primary presenter and contributed major parts of the final presentation slides
  - Co-prepared the proposal presentation slides
  - Assisted in revising and finalizing the final written report
- **Chun-An Chen:**
  - Contributed to the conceptual planning of the Special Case Dataset and assisted in preparing related special-case materials
  - Implemented the initial AUC evaluation pipeline and supported the development of the evaluation framework
  - Primarily drafted the final written report to ensure clarity and consistency, with supportive refinements and contributions from teammates
  - Co-prepared the proposal presentation slides

Appendix: Dataset Details and Additional Samples

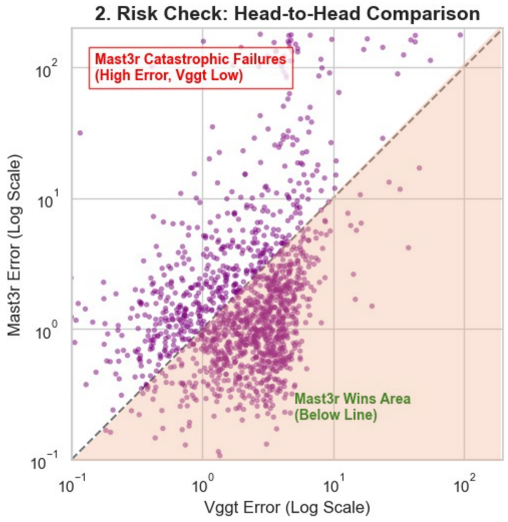

Fig. 12. Head-to-head pose estimation error comparison between MAST3R and VGGT (log scale). Each point corresponds to one image pair under large viewpoint change. The diagonal line denotes equal error; points below the line indicate cases where MAST3R outperforms VGGT, while points above indicate VGGT performs better. Notably, the upper-left region highlights catastrophic failures where MAST3R produces very large errors while VGGT remains stable, revealing a potential reliability risk under challenging conditions.

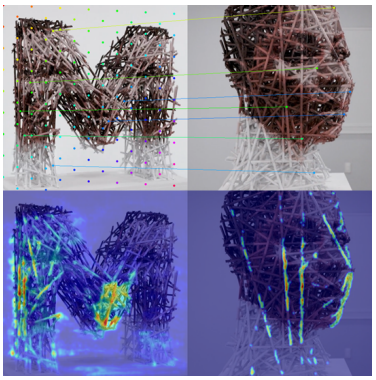

Fig. 13. Perceptual Art: Visualizing the feature matching process and results.

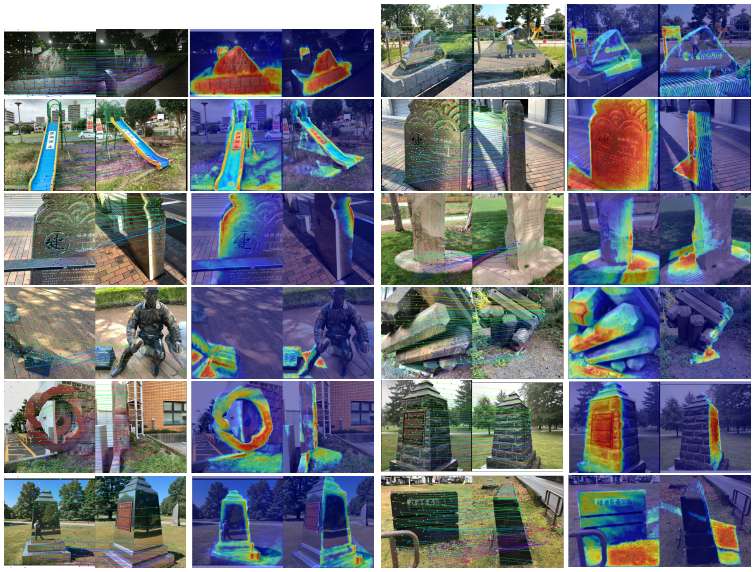

Fig. 14. Additional samples from our dataset showing various scenes and viewpoint changes. Model: VGGT.

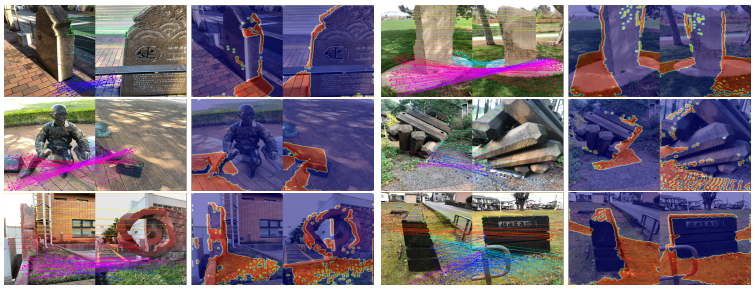

Fig. 15. Additional samples from our dataset showing various scenes and viewpoint changes. Model: MAST3R.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of Large Language Model (LLM) Usage We used multiple Large Language Models, including ChatGPT (GPT-5.2), Claude, and Google Gemini, solely for language-related assistance such as improving writing clarity, refining English grammar, and helping structure paragraphs and summaries. No LLM contributed to experiment design, data analysis, or scientific conclusions. Approximate usage proportions

are: ChatGPT ∼60%, Claude ∼20%, Gemini ∼20%. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final report.

References

1. Leroy, V., Cabon, Y., Revaud, J.: Grounding Image Matching in 3D with MASt3R. In: European Conference on Computer Vision (ECCV). Springer (2024)
2. Wang, J., et al.: VGGT: Visual Geometry Grounded Transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
3. Arnold, E., Wynn, J., Vicente, S., Garcia-Hernando, G., Monszpart, A., Prisacariu, V.A., Turmukhambetov, D., Brachmann, E.: Map-free Visual Relocalization: Metric Pose Relative to a Single Image. In: European Conference on Computer Vision (ECCV). Springer (2022)
4. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D Vision Made Easy. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
5. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning Feature Matching with Graph Neural Networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4938–4947 (2020)
6. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-Free Local Feature Matching with Transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8922–8931 (2021)