# OmniGlue: Generalizable Feature Matching with Foundation Model Guidance
# Supplemental Material

## Appendix

## A. Additional Model Details

OmniGlue undergoes training with $750,000$ iterations using a batch size of $48$ on $8$ NVIDIA Tesla V100 GPUs. The initial learning rate is set at $3e-5$, with a decay rate of $0.999991$ and a hinge step of $55000$. For DINOv2 [9] feature extraction, we use the images with a maximum resolution (long side) of 630, maintaining the aspect ratio during image resizing, for reduce the computation. The DINOv2 backbone employed ViT-14-base [3]. We use the improved positional embedding scheme proposed in LFM-3D [6].

## B. Target Domain Visualization

To illustrate the target image domains we consider in this work, Figure 1 presents example images pairs from each domain, namely: Google Scanned Objects [4], NAVI [5], ScanNet-1500 [1], and DeepAerial [10]. This shows that our target datasets cover a wide range of object and scene types, constituting a challenging task for generalizable image matching.

## C. Area Under Curve (AUC) Pose Results

We also report pose AUC performance, as shown in Table 1. Because the limited performance on out-of-domain data, we report pose accuracy in the main paper.

## D. Latency analysis.

We note that novel OmniGlue modules do not hurt latency as compared the baseline SuperGlue model. Even though DINOv2 introduces additional computation, we use its features to prune the graphs and reduce the computation accordingly.

Theoretically, the computation that DINOv2 introduces is $O(n_1(hw)^2)$, where $n_1 = 9$ (number of DINOv2 attention layers), $h = \frac{H}{14}$ and $w = \frac{W}{14}$ ($H$ and $W$ are input resolution to DINOv2). The computation that pruning saves is $O(2n_2kk^{'})$, where $n_2 = 9$ (number of information propagation blocks), $k = 1024$ (number of target keypoints in one image), $k^{'} = \frac{k}{2}$ (number of pruned keypoints in the other image) and the coefficient 2 is multiplied because there are 2 inter-graph aggregation modules in each block. It is simplified as $O(n_2k^2)$. With the resolution $W = 630$ and a typical aspect ratio of 16:9, the $hw \approx k = 1024$. Thus, the introduced and saved computation are balanced.

We report the empirical speed results in Table 2, which shows that OmniGlue runs at a similar frame rate as the baseline SuperGlue model (no graph pruning). Inference was performed on an NVIDIA A40 GPU with FlashAttention. The result is reproduced with using Glue-Factory.

## E. Additional Qualitative Results

We additionally present qualitative results of OmniGlue in Figure 2. We compare our method (last column) with two reference matching methods: mutual nearest neighbors (MNN, first column) and SuperGlue [11] (second column). We show MNN with SIFT [8] features for two domains, and with SuperPoint [2] features for one. We observe that OmniGlue produces improved matches for image pairs with significant changes in viewing conditions, across a range of domains.

## References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 2

[2] 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*, pages 224–236, 2018. 1, 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Michael Hickman, Krista Reymann, Thomas Barlow McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560, 2022. 1, 2

| | Out-of-domain | | | | |
|---|---|---|---|---|---|
| | *Google Scanned Object* [4] | | *NAVI* [5] | | *ScanNet* [1] |
| | Hard (60-90 degree) | Easy (15-45 degree) | Multiview | Wild | |
| Method | AUC@5°/ 10°/ 20° | AUC@5°/ 10°/ 20° | AUC@5°/ 10°/ 20° | AUC@5°/ 10°/ 20° | AUC@5°/ 10°/ 20° |
| PDCNet [13] | 2.6 / 4.8 / 8.4 | 13.5 / 22.4 / 33.0 | 1.7 / 3.7 / 6.6 | 2.9 / 6.1 / 10.4 | 16.4 / 33.7 / 51.2 |
| LoFTR [12] | 3.6 / 7.3 / 13.0 | 20.7 / 33.9 / 47.9 | 5.7 / 11.8 / 20.4 | 4.5 / 9.4 / 17.0 | 16.9 / 33.6 / 50.6 |
| SIFT [8]+MNN | 3.4 / 6.5 / 11.5 | 16.7 / 30.1 / 40.8 | 3.3 / 6.9 / 12.8 | 2.8 / 5.9 / 11.7 | 1.7 / 4.8 / 10.3 |
| SuperPoint [2]+MNN | 2.5 / 5.3 / 10.0 | 15.2 / 26.1 / 38.8 | 4.5 / 9.7 / 17.8 | 3.7 / 8.0 / 15.1 | 7.7 / 17.8 / 30.6 |
| DINOv2 [9]+SG [11] | 1.8 / 3.6 / 7.4 | 5.5 / 11.6 / 21.3 | 3.3 / 9.7 /.155.6 | 3.8 / 8.4 / 16.3 | 3.3 / 10.0 / 22.0 |
| SuperGlue [11] | 3.4 / 6.9 / 12.2 | 17.5 / 30.1 / 42.6 | 5.1 / 11.2 / 19.9 | 4.8 / 10.2 / 18.3 | 10.4 / 22.9 / 37.2 |
| LightGlue [7] | 3.5 / 7.1 / 12.6 | 18.9 / 32.3 / 46.7 | 5.7 / 12.4 / 21.2 | 4.3 / 9.2 /15.7 | 15.1 / 32.6 / 50.3 |
| **OmniGlue (ours)** | 4.1 / 8.2 / 14.3 | 20.7 / 34.1 / 48.4 | 5.8 / 12.6 / 22.2 | 5.6 / 11.8 / 20.7 | 14.0 / 28.9 / 44.3 |

Table 1. Relative camera pose estimation performance (AUC) and zero-shot generalization capability of models trained on MegaDepth dataset.

| | SuperGlue | OmniGlue |
|---|---|---|
| Speed (FPS) | 52 | 51 |

Table 2. Latency analysis, comparing SuperGlue and our OmniGlue. For both models, we include feature extraction (SuperPoint) and feature matching inference times. Additionally, we include DINOv2 inference time in our measurements for OmniGlue.

[5] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. 1, 2

[6] Arjun Karpur, Guilherme Perrotta, Ricardo Martin-Brualla, Howard Zhou, and Andre Araujo. Lfm-3d: Learnable feature matching across wide baselines using 3d signals. In *Proc. 3DV*, 2024. 1

[7] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *Proc. ICCV*, 2023. 2

[8] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 1, 2

[9] Maxime Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 1, 2

[10] Jae-Hyun Park, Woo-Jeoung Nam, and Seong-Whan Lee. A two-stream symmetric network with bidirectional ensemble for aerial image matching. *Remote Sensing*, 12(3):465, 2020. 1

[11] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 2

[12] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8918–8927, 2021. 2

[13] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5710–5720, 2021. 2

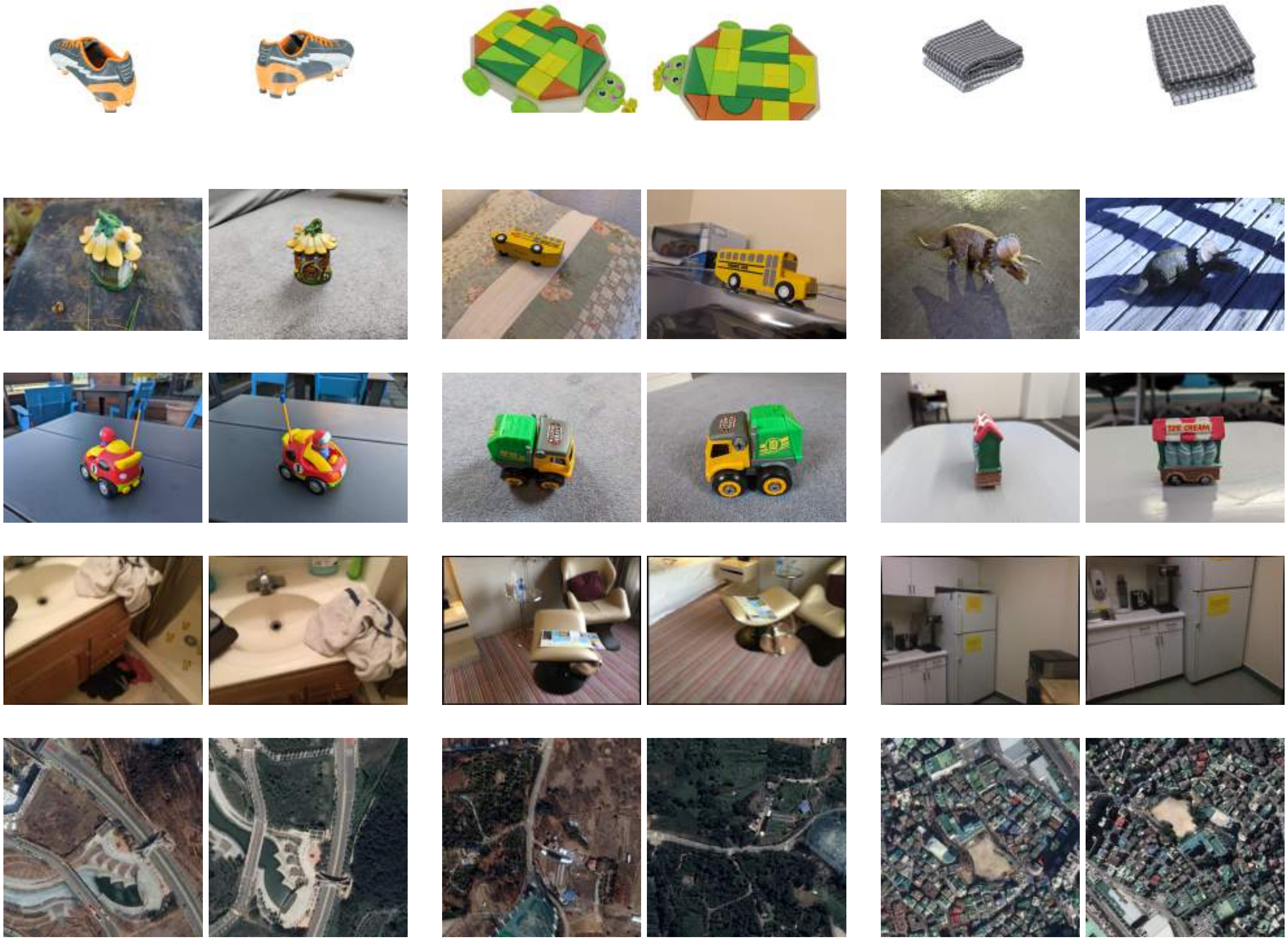

Figure 1. **Target domain examples.** We share some example image pairs from each of the target image datasets. From top row to bottom row, the domains are: Google Scanned Objects (Hard), NAVI Wild Set, NAVI Multiview, ScanNet-1500, and DeepAerial.

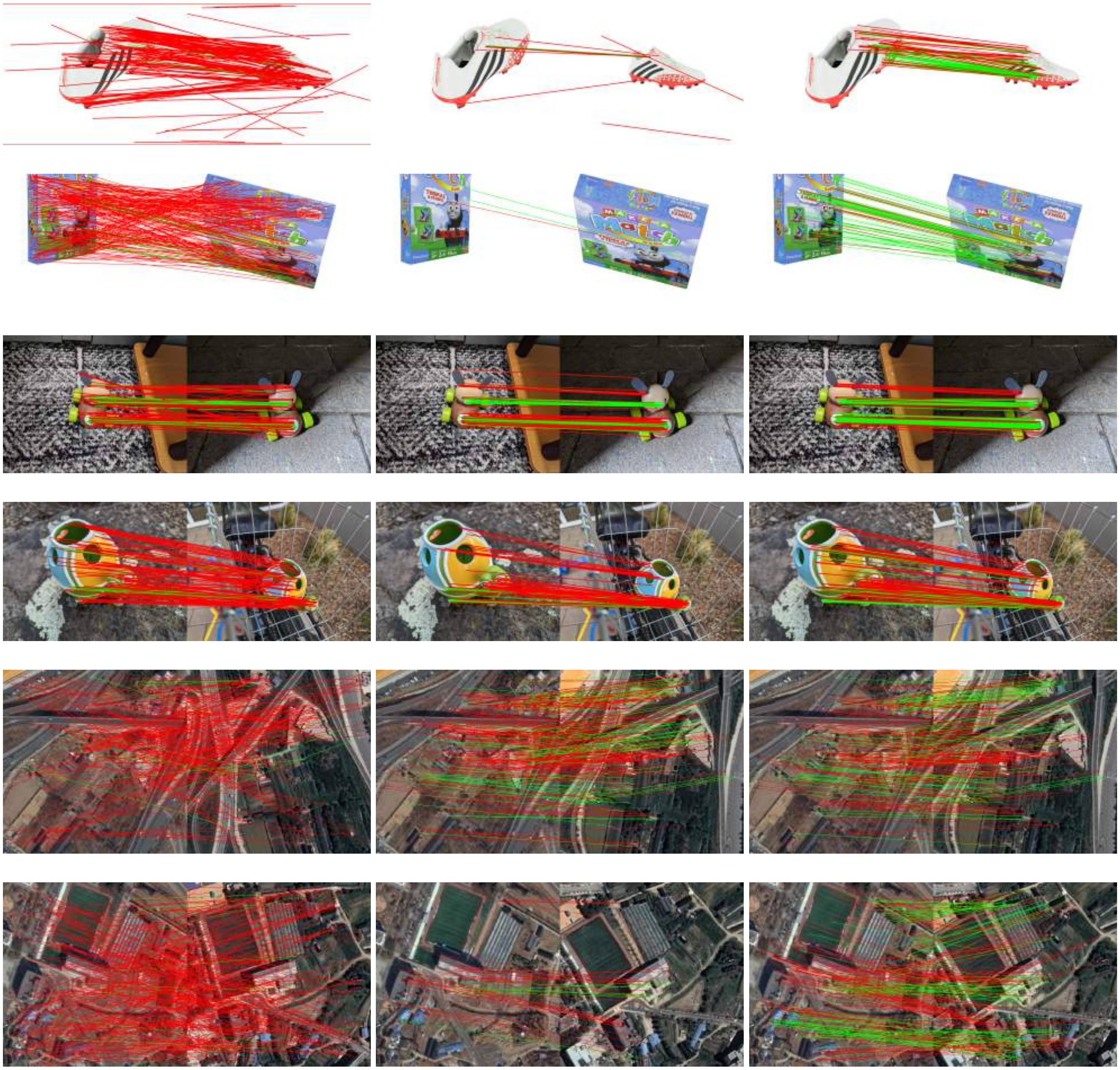

Figure 2. **Qualitative matching comparison.** We compare the following methods: mutual nearest neighbor (MNN, left), SuperGlue (center) and OmniGlue (right). Green lines denote correct correspondences, while red ones denote incorrect predictions. The first two rows present results on Google Scanned Objects (Hard), the following two rows on the NAVI Wild Set, and the final two rows on DeepAerial. The MNN results use SuperPoint features in the first two rows, and SIFT features in the others.