Improving the Ground Truth: MegaDepth in 2023

Alexander Veicht
ETH Zürich
veichta@student.ethz.ch

Andri Horat
ETH Zürich
horatan@student.ethz.ch

Felix Yang
ETH Zürich
fyang@student.ethz.ch

Deep Desai
ETH Zürich
ddesai@student.ethz.ch

Philipp Lindenberger
ETH Zürich
philipp.lindenberger@inf.ethz.ch

Abstract

We present an enhanced version of the MegaDepth dataset, a widely-used collection of unstructured images featuring famous tourist landmarks. MegaDepth has gained significant recognition as a training dataset for tasks such as single-view depth prediction and feature matching. This study proposes an upgraded 3D reconstruction pipeline that integrates state-of-the-art methods at each stage. Additionally, a new image retrieval and model refinement step is introduced to further enhance the overall results as well as a novel dense overlap metrics that can be valuable for training structure-from-motion algorithms. Compared to the original pipeline, our approach achieves a notable increase in the number of registered images and depth values per image.

1. Introduction

The original MegaDepth ground-truth pipeline, described in [8], utilizes structure from motion (SfM) and multi-view stereo (MVS) techniques, alongside data cleaning methods, to generate depth maps and camera poses for images of tourist landmarks. However, this approach exhibits several limitations, including degenerate camera poses, incomplete depth maps, and inaccuracies resulting from unregistered images or noise in the pipeline. These issues primarily arise from the limited robustness of handcrafted features. Despite these drawbacks, MegaDepth has emerged as the industry-standard dataset for training various computer vision models. Its value stems from the dataset's diverse scenes, occlusions, and appearance changes, which enables effective generalization of models [8, 7].

This project introduces a refined MegaDepth ground-truth pipeline that effectively addresses the aforementioned issues through the utilization of deep learning-based methods. Specifically, techniques such as NetVLAD [1], SuperPoint [5], SuperGlue [15], Hierarchical Localization (hloc) [14], Pixel-Perfect Structure-from-Motion (PixSfM) [9], and SegFormer [24] are incorporated. Comparing to the baseline models in the original MegaDepth dataset, our refined pipeline achieves significantly higher precision and detail in reconstructing 3D models of scenes, registering a substantially higher number of images. These improvements naturally result in more accurate and complete depth maps. In addition to providing sparse models, depth maps, and ordinal maps, our dataset offers segmentation maps for all images, along with two newly proposed overlap metrics.

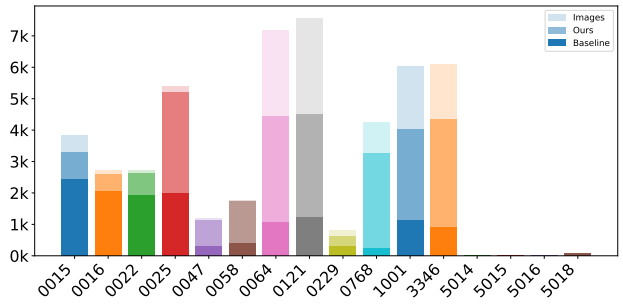


Figure 1: Comparison of registered images between the baseline model (no shade) and our pipeline (light shade). The baseline model achieves approximately 50.3% registered images across the scenes (28.5% when weighted by the number of images), whereas our pipeline achieves a significantly higher registration rate of 86.2% across the scenes (76.3% of all images).

2. Related Work

RGB-D Datasets. Several RGB-D datasets have been developed in the field. ScanNet [4] is an extensive dataset comprising over 2.5 million frames from more than 1500 indoor scenes. Hypersim [13] offers 77,400 images from

461 synthetic indoor scenes created by professional artists. However, both datasets are limited to indoor scenes and have restricted variations in appearance. TartanAir [23] is a synthetic dataset featuring diverse scenes in different environments simulated using Unreal Engine. With 400-3000 data frames per scene, TartanAir captures scenes with various weather conditions, lighting variations, and provides sensor/groundtruth variables. On the other hand, ETH3D [19] and LaMAR [16] focus on high-precision depth maps of real-life locations. However, these datasets lack significant variations in lighting, camera intrinsics, and scenes.

Feature Extraction and Matching. In recent years, machine learning-based approaches have emerged as powerful alternatives to traditional methods for feature extraction and matching. Notably, SuperPoint [5] employs a fully-convolutional architecture and leverages a self-supervised framework using synthetic data to learn keypoint locations and descriptors. Another approach, SuperGlue [15], employs a graph neural network with self- and cross-attention layers to facilitate robust correspondences between local features in two images. LoFTR [20] utilizes transformers for feature matching in a detector-free manner, while DISK [22] adopts reinforcement learning techniques to learn local features via policy gradient methods. These machine learning-based techniques have demonstrated their efficacy in enhancing feature extraction and matching capabilities.

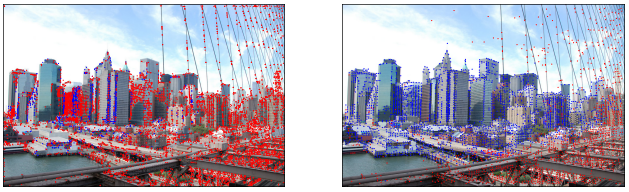
3. Methods

The original ground-truth pipeline relies on SIFT [10] features for matching, using nearest neighbors of each SIFT descriptor. The reconstruction of sparse and dense models is performed using COLMAP [17, 18]. Post-processing steps are employed to refine the obtained depth maps from multi-view stereo. We refer to the MegaDepth paper [8] for a more detailed overview of their pipeline.

In our refined pipeline, an image retrieval step is introduced to enhance runtime efficiency when utilizing deep learning-based extractors and matchers. Following feature matching and extraction, a 3D model is reconstructed using the standard SfM library in COLMAP with PixSfM being added to improve the quality of the sparse model. Depth maps are then generated using the standard MVS approach in COLMAP. Finally, a post-processing step is implemented to clean the depth maps and obtain additional ordinal maps.

3.1. Image Retrieval

For scenes with a large number of images, it becomes infeasible to consider every possible image pair for matching, especially with slow deep-learning matchers. To address this limitation, image retrieval methods are employed to select the most similar images given a query image. This helps to reduce the runtime from quadratic to merely linear in the number of images.



(a) SIFT features (b) SuperPoint features

Figure 2: Comparison of feature points between SIFT and SuperPoint in an image. The feature points are colored based on their visibility, where blue represents triangulated points and red represents non-triangulated points. SuperPoint exhibits a significantly higher number of visible points compared to the SIFT model. Although SuperPoint may have some points in the sky, they are effectively discarded during the SfM.

For this purpose, global features are extracted using either NetVLAD [1] or CosPlace [3]. The selection of image pairs to be considered involves choosing the top 50 images based on the similarity of their global features to the query image.

3.2. Feature Extraction and Matching

The primary alternative explored for SIFT was SuperPoint [5]. SuperGlue [15] was selected as the feature matcher, as it is specifically optimized for SuperPoint and known for its robustness. Additionally, experiments were conducted with dense matching techniques, specifically LoFTR [20]. Evaluation of the various retrieval, extractor, and matcher combinations was performed. The results are shown in Table 1.

In comparison to SIFT, significant changes were observed in the number of observations and registered images. However, it should be noted that the learning-based extractor exhibited a higher mean reprojection error when compared to SIFT. This discrepancy can be attributed to SIFT's superior localization and sub-pixel accuracy.

3.3. Refinement

An additional step was incorporated into the ground-truth pipeline to enhance the sparse model through the utilization of Pixel-Perfect Structure-from-Motion (PixSfM) [9]. PixSfM employs off-the-shelf CNNs to extract dense features. These dense features are then utilized in a series of refinements within PixSfM. These refinements include a keypoint adjustment step, which aims to enhance sparse features by increasing track length, as well as bundle adjustment after Structure-from-Motion (SfM) using a featuremetric error.

3.4. Post-processing

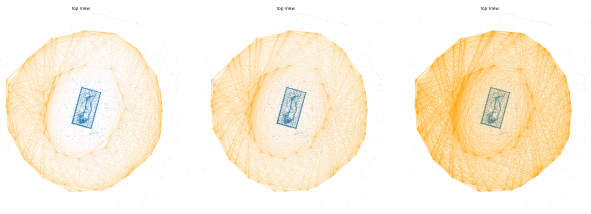
For the final post-processing stage of the pipeline, the existing approach from the original pipeline was primarily adopted by re-implementing their MATLAB code in Python. Various steps were performed, including the application of median filtering to filter out unstable depth values, the removal of small connected components, and the utilization of morphological operations. Additionally, semantic segmentation models fine-tuned on the ADE20K [25] dataset were employed to eliminate depth values from the sky region and various foreground objects such as humans or cars. Three model architectures, namely HRNetV2 [21], BEiT [2], and SegFormer [24], were investigated for this purpose.

In addition to the depth maps, ordinal maps are computed for each image in the pipeline. The computation of these ordinal maps involves selecting appropriate foreground and background regions based on the segmentation maps and post-processed depth maps. Connected components from the segmentation maps are considered if they are sufficiently large and belong to a predefined foreground or background class. For background components, an additional check is performed to ensure that the corresponding depth values from the depth maps fall within the last quartile of all valid depth values, thereby ensuring the correctness of the ordinal relations. Such ordinal maps provide additional depth information that can be leveraged during the training of a depth prediction model, as most foreground objects do not contain valid depth values and are missing from the post-processed depth maps.

3.5. Overlap Metric

Prior studies [6] have employed sparse point cloud overlap to identify suitable image pairs for training keypoint extractors. The sparse overlap score is determined by the ratio of matched keypoints across the pair to the total number of keypoints in the first image. This overlap score serves as a valuable indicator of the image pair's difficulty level. However, there are two significant drawbacks to this approach. Firstly, unmatched keypoints introduce a bias towards a score of zero. Secondly, an uneven distribution of keypoints results in a noisy score. To address these issues, a robust overlap metric is necessary to generate appropriate training data for keypoint extractors and matchers, ensuring the desired level of difficulty.

Two novel overlap scores are introduced in our proposal, taking into account the area of overlap and changes in the view direction. The first score, referred to as dense overlap score, quantifies the overlap area by considering the number of depth values in the first image that align with the second image, divided by the total number of valid depth values in the first image. The second score, the cosine-weighted dense overlap, incorporates the cosine of the view directions for each overlapping point. This weighting penalizes slanted surfaces and provides a more accurate assessment of the overlap. Qualitative comparisons are shown in Figure 3 and Figure 4.



Sparse Dense cos-weighted Dense

Figure 3: The image showcases a top-down perspective of a sparse reconstruction depicted in blue, with camera poses visualized in orange. The intensity of the color represents the degree of overlap between two images, with higher opacity indicating a higher overlap.

The sparse overlap of an image pair indicates a high level of shared triangulated points between the two images. On the other hand, the dense overlap measures the extent to which the images share a large area in the dense reconstruction. In contrast, the cosine-weighted dense overlap score provides a more realistic assessment by taking into account the viewing angle to the surface. This score is inspired by Lambert's cosine law, which reduces the intensity of a ray based on the cosine of the emission angle. For each overlapping point in the dense reconstruction, the cosine of the viewing angle in both images is considered, and the minimum of these two values is used as a weight when summing over all shared points. The minimum is chosen as it represents the more challenging scenario. It's worth noting that the cosine-weighted overlap between two identical images results in the average cosine of the viewing angle, and both the sparse and dense overlap scores yield a value of 1 in such cases.

4. Results and Discussion

The implementation of the proposed pipeline resulted in reconstructed sparse models that showcased substantial improvements in both detail and completeness compared to the baseline models (Figure 5). Additionally, there was a notable increase in the number of observations and registered images within our models (Figure 1). The enhancements in camera poses were particularly evident in Figure 6, where the improved alignment can be attributed to the camera poses, as both scenes had the same number of registered images.

The post-processed depth maps generally exhibited satisfactory quality. However, certain instances revealed unsightly artifacts attributed to the diverse range of image

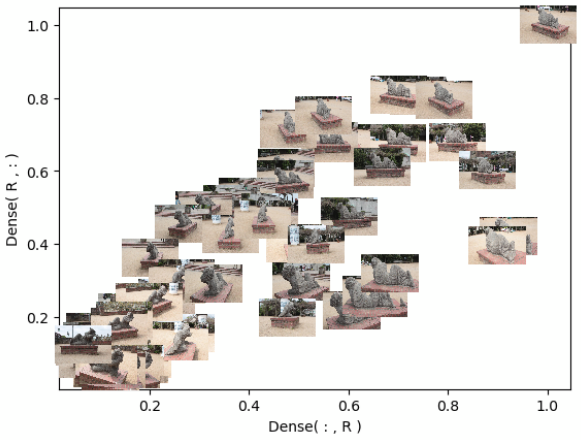


Figure 4: The figure illustrates the dense overlap metric, with images positioned on the x-axis based on their overlap with the anchor image in the top right corner. Along the y-axis, the positioning reflects the overlap from the anchor to each respective image.

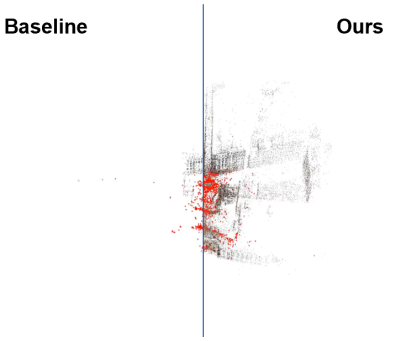


Figure 5: Comparison between the baseline sparse model (left) and our enhanced sparse models (right) for Piccadilly Circus, London (0016). Our models exhibit a substantially higher number of points, attributed to the increased count of registered images and observations per image.

sizes. These artifacts arose from the utilization of fixed kernel sizes for filtering and morphological operations. Upon visual examination of the segmentation and depth maps generated by HRNetV2, BEiT, and SegFormer across multiple scenes, SegFormer consistently demonstrated superior accuracy and robustness. Hence, SegFormer was chosen as the ultimate segmentation model, as illustrated in Figure 7.

Concerning image retrieval, the restriction on the number of images employed in matching adversely affected several metrics, including the mean reprojection error, the number of registered images, and the total number of observations. However, image retrieval is crucial when working

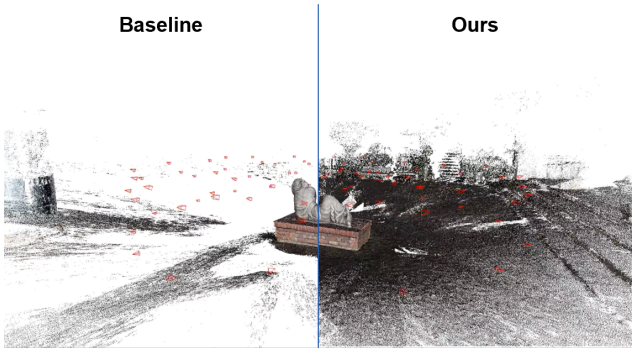


Figure 6: Comparison of dense models of Colony on Mathildenhöhe, Darmstadt (5018), highlighting the impact of improved camera poses on both models with an equal number of images.

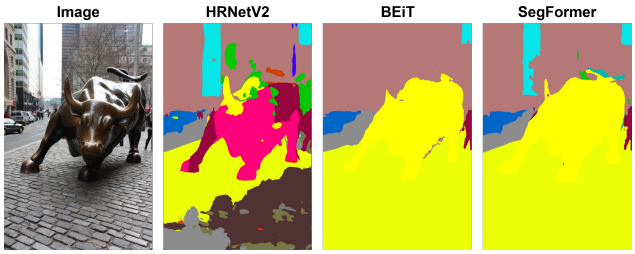


Figure 7: A comparison between the three different segmentation models.

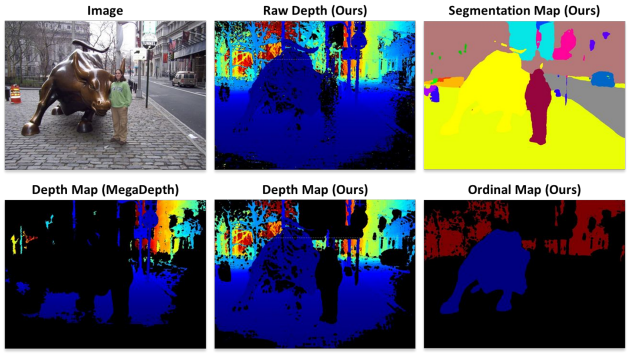


Figure 8: Showcase of post-processing steps. Note how the depth values of the person, car and other foreground objects are removed using the segmentation map.

with learning-based extractors and matchers, as they tend to exhibit slower performance. Through our experimentation, we identified a value of 50 pairs per image as optimal, striking a favorable equilibrium between our metrics and reasonable processing time.

In addition to the primary improvement of achieving a higher number of registered images and more observations per registered image, our pipeline exhibits a notable in-

Name	Registered	Rep. Error
Baseline	318	**0.59**
SIFT+NN (netvlad-5)	359	0.84
SIFT+NN (netvlad-50)	561	0.96
SIFT+NN (cosplace-50)	558	0.95
SIFT+NN (exhaustive)	635	0.88
LoFTR (netvlad-50)	501	1.24
SP+SG (netvlad-50)	676	1.39
SP+SG (cosplace-50)	**679**	1.38

Table 1: Summary of different configurations for Charging Bull, New York (0229). Increasing the number of retrievals and incorporating SuperPoint with SuperGlue led to a noticeable increase in the number of registered images. Using CosPlace over NetVLAD had no big effect.

crease in the inclusion of nighttime images, a few examples are shown in Figure 9. This augmentation contributes to a more robust training dataset. Furthermore, our analysis of unregistered images revealed a consistent pattern in their failure to register, reinforcing our confidence in the overall completeness of the pipeline. Many of these unregistered images comprised unrelated content, such as images capturing nearby basketball matches or fruits in a local market. Additionally, a small fraction of unregistered images exhibited filters or contained spurious objects, such as crowds of humans in protest scenarios. Notably, the number of unregistered rotated images was minimal compared to the overall count of unregistered images.

Figure 9: Some examples of registered nighttime images with overlaid depth maps. These images are not registered in the baseline models.

Lastly, an anticipation of symmetry-related challenges arose in relation to SuperPoint and SuperGlue. To investigate this concern, we hand-picked scenes from the dataset that appeared to exhibit the highest degree of symmetry. Surprisingly, we did not encounter any notable issues with symmetry in most cases. However, a specific phenomenon came to our attention in one particular scene—the Tower of London. The sparse 3D reconstruction of this scene revealed the presence of "ghost towers," appearing as floating structures alongside the original towers, as depicted in Figure 10. Our leading hypothesis attributes this phenomenon to a limitation in image retrieval, specifically the absence of matches between images captured in close proximity to the tower and those taken from a greater distance. This constraint presents a significant obstacle in generating accurate representations and depth maps for this scene.

Figure 10: A depiction of the ghost tower phenomenon.

5. Image Matching Challenge 2023

Additionally, participation in the Image Matching Challenge 2023 allowed us to leverage our expanded knowledge of 3D reconstruction. As of the present writing, our standing in the competition is 5th place. This experience yielded several valuable insights that we can apply to enhance our pipeline:

- The limitations of rotation invariance in deep features became evident. While rotation invariance is a desirable trait, achieving it often comes at the expense of specificity. Consequently, registering rotated images posed a challenge as keypoint correspondences between upright and rotated images were scarce. To address this issue, we incorporated a vision transformer [11] capable of predicting the orientation of an image. This enabled us to rotate each image by multiples of 90 degrees, if necessary.
- We significantly improved our competition score by using ensembles of feature extractors and matchers. Combining deep features with conventional features such as SIFT turned out to be very powerful as SIFT features are invariant to scale and rotation.
- For small scenes with each image of the same size, forcing each image to have the same camera intrinsics resulted in an improvement in camera poses. While this proved invaluable in the challenge, the MegaDepth dataset contains very few such scenes, and thus this technique may not be possible in our pipeline.

We believe that many of the above insights can help us with the discussed problems. Specifically, employing an ensemble of our current feature extractor alongside SIFT, combined with exhaustive retrieval, can provide sufficient information to eliminate the ghost tower phenomenon.

6. Future Work

While the results obtained demonstrate promise in establishing a new training dataset standard, there are opportunities for improvement. Specifically, our pipeline will benefit from incorporating the insights gained from the Image Matching Challenge 2023, enhancing the 3D reconstruction process and achieving more accurate camera poses. Additionally, careful consideration is required for the post-processing of raw depth maps, as this step significantly impacts the quality of the final depth maps. Future work might also explore combining MVS and NeRF's [12] to obtain accurate and more dense depth maps. Expanding the dataset to include additional landmarks and diverse scene types is also under consideration, aiming to create a more varied and comprehensive dataset. Furthermore, aligning our models with street maps would enable metric values for the depth maps. With these refinements, we aim to create a robust pipeline and generate a ground truth dataset that contributes significantly to the computer vision community.

7. Contributions

Alexander Veicht: Implemented the pipeline, wrote configs for all different feature extractors and matchers, created fly-through visualization movies. Conducted experiments and analyses. Created ensemble method used in the image matching challenge.

Andri Horat: Implemented model alignment with the ground orientation using PCA, cosine-weighted overlap & visualizations. Conducted experiments and analyses. Setup orientation prediction model for the image matching challenge.

Felix Yang: Implemented the dense overlap score, semantic segmentation and depth cleaning. Contributed to the pipeline. Conducted experiments and analyses. Experimented with cropping ensembles for the image matching challenge.

Deep Desai: Implemented initial version of dense overlap scores. Investigated NeRFs. Contributed to the pipeline. Conducted experiments and analyses, especially regarding retrieval. Built wheels for the image matching challenge.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 4321, 4322

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 4323

[3] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022. 4322

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 4321

[5] 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. 4321, 4322

[6] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019. 4323

[7] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 4321

[8] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 4321, 4322

[9] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5987–5997, 2021. 4321, 4322

[10] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 4322

[11] Subhadip Maji and Smarajit Bose. Deep image orientation angle detection. *arXiv preprint arXiv:2007.06709*, 2020. 4325

[12] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 4326

[13] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 4321

[14] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical

localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 4321

[15] 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. 4321, 4322

[16] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L. Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. LaMAR: Benchmarking Localization and Mapping for Augmented Reality. In *ECCV*, 2022. 4322

[17] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4322

[18] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 4322

[19] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4322

[20] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 4322

[21] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 4323

[22] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 4322

[23] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 4322

[24] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 4321, 4323

[25] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 4323