MuM: Multi-View Masked Image Modeling for 3D Vision

Supplementary Material

A. Further Details on Pretraining

Data sequences. Below, we outline the selection of sequences of frames. In the following section, we detail how to sample batches from sequences. For information on the dataset mixture, we refer to Section 3.4 of the main manuscript. Our protocol for finding sequences is simple, as strict co-visibility is not required. This greatly simplifies things and allows us to, without much effort, scale to around 20 million individual frames. MegaDepth and MegaScenes are the only datasets where we employ a more complicated sampling scheme. This is because each scene of the dataset is highly diverse and has no natural temporal sequencing.

For the former, we use the image overlap scores computed by D2-Net [22] and greedily sample 1,000 sequences per scene by randomly selecting an anchor frame and then chaining together images based on positive overlap. For the latter, we compute the IoU heuristic similar to CroCo and iteratively select the next image that maximizes 3D point overlap with the current image while maintaining sufficient camera viewpoint diversity. Starting from images with the most visible 3D points, we greedily build sequences until no suitable next image can be found, discarding sequences shorter than 24.

For datasets where each scene is small (*e.g.*, CO3D, DL3DV10K, Hypersim, WildRGBD), we simply select all frames in the scene as sequences and randomly sample from those during training. For the remaining datasets (3DStreetView, RealEstate10K, ScanNet++, BlendedMVS, and ARKitScenes), we use simple heuristics to form sequences. Since the frames per scene are numerically ordered, where adjacent frames are also spatially adjacent in the scene, we chunk them. We find that a chunk size of 100 frames usually works fine. After forming sequences of frames, we now detail how to select the frames in a batch.

Sampling. For each batch, independently on each GPU, we either (i) sample a full batch from ImageNet-1K (10% probability) or (ii) sample multi-view data (90% probability). To sample multi-view data, we uniformly sample a sequence length S between 2 and 24 frames. Given the sequence length, we then iteratively sample datasets with the weighting given in Table 1 and uniformly select a sequence therein. See Figure 5 for qualitative examples of sequences. We stop when the total number of frames per GPU is about to exceed 96. This gives us a batch shape of $(\text{floor}(\frac{96}{S}), S, 3, 256, 256)$. This is done independently per GPU. We sometimes refer to 96 as the batch size and use this in the linear scaling rule for learning rate [29].

Figure 5. **Data examples**. We visualize random sampled sequences from the dataset. Sequence lengths vary.

Architecture. The total model is made up of a ViT-L encoder (width 1024, depth 24 and 16 heads) followed by a ViT-B multi-view decoder (width 768, depth 12 and 12 heads). This is the exact same size as the final CroCo v2 model and thus also DUSt3R and MASt3R (excluding heads). Half of the Transformer blocks in the multi-view decoder have a global attention window and half only attend to patches within the same frame. The first block is a frame-wise attention block. The total number of trainable parameters in the model is 389.5 million, whereof around 302.3 million come from the encoder. The final layer, after the multi-view decoder, is a linear layer $W \in \mathbb{R}^{768\times768}$. The dimensions can be understood from the way that it maps each patch between the 768 embedding dimensions of the Transformer decoder to the $3 \times 16 \times 16 = 768$ color values in a patch. The loss is then computed by comparing the output of this linear mapping to the normalized (mean subtracted and divided by the standard deviation) pixel values in the ground truth patch.

Compute Resources. In Table 12, we detail the computing resources used for the main experiments. Failed or discarded training runs are excluded. The evaluations not included required only a linear probe or a k-NN classifier. These were run on a single A100 and took no longer than a couple of hours. The reporting is conducted towards academic transparency in the compute resources required to replicate our results.

B. Further Details on Experiments

Feedforward 3D reconstruction. We take inspiration from the design of VGGT and use the same heads, excluding the tracking head. Namely, we use DPT heads for world points and depths and an iterative ViT for the camera head.

Table 12. **GPU hours.** We report the compute used for the main experiments. We omit linear probing and k-NN validations.

Experiment	Model	GPUs	Time
Pretraining	MuM[†]	32×A100	3 days
Pretraining	MuM	64×A100	3 days
RoMa	MuM	4×A100	5 days
RoMa	DINOv3	4×A100	5 days
RoMa	CroCov2	4×A100	5 days
Rel. pose	MuM	32×A100	1.5 days
Rel. pose	DINOv3	32×A100	1.5 days
Rel. pose	CroCov2	32×A100	1.5 days
3D recon. (frozen)	MuM	32×H200	3 days
3D recon. (frozen)	DINOv3	32×H200	3 days
3D recon. (frozen)	CroCov2	32×H200	3 days
3D recon. (finetune)	MuM	16×H200	3 days
3D recon. (finetune)	Rand.	16×H200	3 days

[†]A lighter training run using 66 A100 days vs. 96 of CroCov2.

For the frozen encoder experiments, the points and depth heads also output confidence maps, we train for 100K steps with a learning rate of 2×10^{-4} on 32×H200 for around 3 days, and sample data following VGGT, namely, using 48 frames per GPU and an image resolution of 512. We use the AdamW [40] optimizer and minimize the same training loss as in VGGT:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{camera}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{pmap}} \tag{3}$$

Where the camera loss is given by $\mathcal{L}_{\text{camera}} = \sum_{i=1}^{N} ||\hat{c}_i - c_i||_{\mathcal{H}}$ where c_i are the ground truth camera parameters and $\hat{c}_i$ are the predictions in frame i and $|| \cdot ||_{\mathcal{H}}$ denotes the Huber loss. The depth objective, $\mathcal{L}_{\text{depth}}$, follows DUSt3R and uses an aleatoric-uncertainty loss. We further incorporate the gradient-based term used in VGGT. The pointmap loss, $\mathcal{L}_{\text{pmap}}$, is similarly defined.

The finetuning distillation protocol is slightly lighter using half the batch size (due to training on half the number of GPUs) and with a learning rate of 1×10^{-4}. The distillation loss is simply the sum over the L2 errors (given in Equation 2). We train without confidence weights.

Dense matching. For the linear probing evaluation we take as basis the MegaDepth1500 and ScanNet1500 dense evaluation protocols in RoMa and DKM. We train a square linear projection on top of the frozen encoder features followed by a log-softmax kernel nearest neighbor matcher to estimate the dense warp. We achieve a robust metric by sweeping over learning rates and weight decays in parallel. We train for 25K steps with batch size 8 and evaluate every 2,500 steps, reporting the best performance achieved. The evaluation protocol takes roughly an hour on an A100.

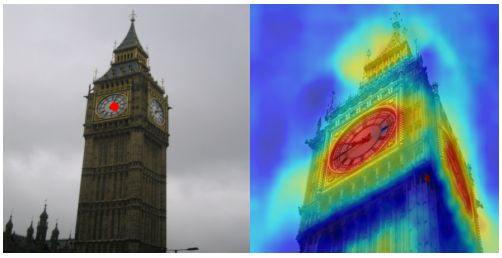

Figure 6. **Attention map for query patch**. We visualize the global attention map in the decoder for a query patch. We find that the attention score is highest for matching patches.

For RoMa, we follow their exact training setup. which is the following. Training for 3 million steps on 560×560 resolution with a frozen encoder. This takes roughly 5 days on $4\times$A100-80GB. We opt to only train on MegaDepth (commonly referred to as outdoor) and use this model for all evaluations.

Two-view relative pose estimation. The decoder architecture follows exactly Reloc3r [20] but is not initialized from DUSt3R but instead trained from scratch. We train the decoder and the heads with a learning rate of 2×10^{-4} and finetune the encoder with a learning rate of 2×10^{-5}. We use a total batch size of 1024 and train on 16×A100 for approximately 1 day.

Monocular depth estimation. We follow the protocol in DINOv2 [43]. Namely, we use the MMCV [15] package and apply random cropping, rotation and flipping augmentation. We crop the images to 480×640 and treat it as a classification problem with 256 bins. We use a combination of cross entropy and gradient loss. We train with AdamW, cosine annealing and a peak learning rate of 1×10^{-4}.

Semantic recognition. For both classification and segmentation, we follow exactly the protocol in CAPI [17]. In particular, the classification protocol uses an attentive probe. This is due to MuM, CroCo v2 and MAE, not having a meaningful global representation in the same way as the DINO family (following its instance discrimination objective on the CLS token). The attentive probe is trained in a supervised fashion and its output is processed by a trained linear layer.

C. Additional experiments

Qualitative examples. Inspired by ZeroCo [1], we visualize the global attention map between a query patch in the first image and all the patches in the second image. The result, illustrated in Figure 6, clearly illustrates that the attention map contains information about correspondences be-

Table 13. **Cosine similarity feature matching**. We report average recall on NAVI [33] following the protocol of Probe3D [26] and EPE on MegaDepth and ScanNet.

Method	NAVI ↑	MegaDepth ↓	ScanNet ↓
MAE	43.7	88.8	83.4
CroCo v2	41.0	60.3	62.1
DINOv3	**62.8**	**36.5**	56.4
MuM	46.9	50.8	**55.7**

tween the two frames. An interesting future direction is to investigate zero-shot matching directly from the attention map, similar to ZeroCo [1].

We also visualize the dense warp estimated through linear probing in Section 4.2 in Figure 7.

Cosine similarity feature matching. In our main results, we report the linear probing dense feature matching accuracy. However, prior work has explored the matchability of the raw features. While El Banani et al. [26] notes that more expressive probes are necessary to determine 3D understanding, they opt for correspondences based on cosine similarity in their paper. We follow their protocol and report the matching performance in Table 13. As suggested by Siméoni et al. [53], we evaluate both with and without layer normalization on the frozen features and report the best result. Interestingly, generative objectives such as MuM, MAE, and CroCo v2 perform worse when evaluated without a linear probe. This might be related to how the DINO-family uses a patch similarity objective whereas the generative objectives tested work in pixel space. We further find that MuM outperforms CroCo v2 and MAE and matches DINOv3 on ScanNet.

Figure 7. **Dense warp**. Warp estimated through linear probing on MegaDepth on top of frozen features.

References

[1] Honggyu An, Jin Hyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hong, and Seungryong Kim. Cross-view completion models are zero-shot correspondence estimators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1103–1115, 2025. 2, 3

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023. 6

[3] Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 6

[4] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision – ECCV 2022*, pages 348–367, Cham, 2022. Springer Nature Switzerland. 2

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 3

[6] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Con-*

ference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021. 4

[7] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv:2504.13181*, 2025. 6

[8] Lucas Brynte, José Iglesias, Carl Olsson, and Fredrik Kahl. Learning structure-from-motion with graph attention networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[9] Georg Bökman, David Nordström, and Fredrik Kahl. Flopping for FLOPs: Leveraging equivariance for computational efficiency. In *Forty-second International Conference on Machine Learning*, 2025. 8

[10] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 5

[11] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1050–1060, 2025. 2

[12] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1050–1060, 2025. 2

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 1, 2

[14] João Carreira, Dilara Gokay, Michael King, Chuhan Zhang, Ignacio Rocco, Aravindh Mahendran, Thomas Albert Keck, Joseph Heyward, Skanda Koppula, Etienne Pot, Goker Erdogan, Yana Hasson, Yi Yang, Klaus Greff, Guillaume Le Moing, Sjoerd van Steenkiste, Daniel Zoran, Drew A. Hudson, Pedro Vélez, Luisa Polanía, Luke Friedman, Chris Duvarney, Ross Goroshin, Kelsey Allen, Jacob Walker, Rishabh Kabra, Eric Aboussouan, Jennifer Sun, Thomas Kipf, Carl Doersch, Viorica Pătrăucean, Dima Damen, Pauline Luc, Mehdi S. M. Sajjadi, and Andrew Zisserman. Scaling 4d representations, 2025. 2

[15] MMCV Contributors. MMCV: OpenMMLab computer vision foundation. https://github.com/open-mmlab/mmcv, 2018. 2

[16] 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. 5, 6

[17] Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latent patches for improved masked image modeling. *Transactions on Machine Learning Research*, 2025. 2, 3, 6

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 2

[20] Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16739–16752, 2025. 2, 6

[21] 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 3

[22] 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. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

[23] 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 (CVPR)*, pages 17765–17775, 2023. 2

[24] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 4, 5

[25] Johan Edstedt, David Nordström, Yushan Zhang, Georg Bökman, Jonathan Astermark, Viktor Larsson, Anders Heyden, Fredrik Kahl, Mårten Wadenbäck, and Michael Felsberg. Roma v2: Harder better faster denser feature matching, 2025. 8

[26] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21795–21806, 2024. 6, 7, 3

[27] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pre-

training of large autoregressive image models. In *Int. Conf. Machine Learning*, 2024. 2

[28] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, 2023. 3

[29] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2018. 4, 1

[30] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004. 2

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 1, 2, 6

[32] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[33] 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. In *NeurIPS*, 2023. 3

[34] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanaes. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5

[35] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*. IEEE, 2026. 2

[36] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 4, 5, 6

[37] 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 (ICCV)*, pages 17627–17638, 2023. 2

[38] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 4

[39] Thibaut Loiseau, Guillaume Bourmaud, and Vincent Lepetit. Alligat0r: Pre-training through covisibility segmentation for relative camera pose regression. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 3, 2

[41] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 6

[42] David Nordström, Johan Edstedt, Fredrik Kahl, and Georg Bökman. Octic vision transformers: Quicker vits through equivariance, 2025. 8

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 1, 2, 6

[44] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. Accessed: 2024-11-15. 2

[45] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 5

[46] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 4, 5

[47] 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 *International Conference on Computer Vision (ICCV) 2021*, 2021. 4, 5

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 4

[49] 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. 6

[50] Thomas Schops, Johannes L. Schonberger, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[51] Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multi-view masked world models for visual robotic manipulation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 30613–30632. PMLR, 2023. 2

[52] Ketul Shah, Robert Crandall, Jie Xu, Peng Zhou, Marian George, Mayank Bansal, and Rama Chellappa. Mv2mae: Multi-view video masked autoencoders, 2024. 2

[53] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 1, 2, 3, 4, 6

[54] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. 3

[55] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *CoRR*, abs/2104.00680, 2021. 2, 6

[56] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 6

[57] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5283–5293, 2025. 2

[58] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[59] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 2

[60] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 2

[61] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 6

[62] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *ECCV*, 2024. 4

[63] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 6

[64] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory, 2024. 2

[65] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer, 2025. 1, 2, 5, 6

[66] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023. 2

[67] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2, 4, 6

[68] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *The Fourteenth International Conference on Learning Representations*, 2026. 3

[69] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 1, 2

[70] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 2, 6

[71] Ziyang Wu, Jingyuan Zhang, Druv Pai, Xudong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma. Simplifying dino via coding rate regularization. In *Int. Conf. Machine Learning*, 2025. 7

[72] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4, 5

[73] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 2

[74] Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21924–21935, 2025. 2

[75] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 4, 5

[76] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 4

[77] Vlardimir Yugay, Duy-Kien Nguyen, Theo Gevers, Cees G. M. Snoek, and Martin R. Oswald. Visual odometry with transformers, 2025. 2

[78] Amir R Zamir, Tilman Wekel, Pulkit Agrawal, Colin Wei, Jitendra Malik, and Silvio Savarese. Generic 3D representation via pose estimation and matching. In *European Conference on Computer Vision*, pages 535–553. Springer, 2016. 4

[79] Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, Sebastian Scherer, and Wenshan Wang. Ufm: A simple path towards unified dense correspondence with flow. In *neurips*, 2025. 4

[80] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023. 5

[81] 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*, 2017. 6

[82] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 6

[83] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022. 2, 6

[84] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. 4, 5